Document made available under the Patent Cooperation Treaty (PCT)

International application number: PCT/US05/007249

International filing date: 03 March 2005 (03.03.2005)

Document type: Certified copy of priority document

Document details: Country/Office: US
Number: 60/550,074
Filing date: 04 March 2004 (04.03.2004)

Date of receipt at the International Bureau: 25 April 2005 (25.04.2005)

Remark: Priority document submitted or transmitted to the International Bureau in compliance with Rule 17.1(a) or (b)

1309829

THE UNITED STATES OF AMERICA

TO ALL TO WHOM THESE PRESENTS SHALL COME:

UNITED STATES DEPARTMENT OF COMMERCE

United States Patent and Trademark Office

April 18, 2005

THIS IS TO CERTIFY THAT ANNEXED HERETO IS A TRUE COPY FROM THE RECORDS OF THE UNITED STATES PATENT AND TRADEMARK OFFICE OF THOSE PAPERS OF THE BELOW IDENTIFIED PATENT APPLICATION THAT MET THE REQUIREMENTS TO BE GRANTED A FILING DATE.

APPLICATION NUMBER: ***60/550,074***
FILING DATE: ***March 04, 2004***
RELATED PCT APPLICATION NUMBER: ***PCT/US05/07249***

UNITED STATES PATENT AND TRADEMARK OFFICE
DEPARTMENT OF COMMERCE

Certified by

Jon W. Dudas

Under Secretary of Commerce
for Intellectual Property
and Director of the United States
Patent and Trademark Office

PTO/SB/16 (08-03)
Approved for use through 7/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

PROVISIONAL APPLICATION FOR PATENT COVER SHEET

This is a request for filing a PROVISIONAL APPLICATION FOR PATENT under 37 CFR 1.53(c).

Express Mail Label No. EJ624594035US

INVENTOR(S)

Given Name (first and middle [if any])	Family Name or Surname	Residence (City and either State or Foreign Country)
Christopher T.	Harbison	Quincy, Massachusetts
David B.	Gordon	Somerville, Massachusetts
Richard A.	Young	Weston, Massachusetts

Additional inventors are being named on the ______ *separately numbered sheets attached hereto*

TITLE OF THE INVENTION (500 characters max)

TRANSCRIPTIONAL REGULATORY CODES OF EUKARYOTIC GENOMES AND METHODS THEREOF

Direct all correspondence to: **CORRESPONDENCE ADDRESS**

[x] Customer Number: 28120

OR

[ ] Firm or Individual Name	Ropes & Gray LLP				
Address	Patent Group One International Place				
City	Boston	State	MA	Zip	02110
Country	US	Telephone	617-951-7000	Fax	617-951-7050

ENCLOSED APPLICATION PARTS (*check all that apply*)

[x] Specification *Number of Pages* 37
[x] Drawing(s) *Number of Sheets* 39
[x] Application Data Sheet. See 37 CFR 1.76
[ ] CD(s), Number
[ ] Other (specify):

METHOD OF PAYMENT OF FILING FEES FOR THIS PROVISIONAL APPLICATION FOR PATENT

[ ] Applicant claims small entity status. See 37 CFR 1.27.
[ ] A check or money order is enclosed to cover the filing fees.
[x] The Director is hereby authorized to charge filing fees or credit any overpayment to Deposit Account Number: 18-1945
[ ] Payment by credit card. Form PTO-2038 is attached.

FILING FEE AMOUNT ($): 160.00

The invention was made by an agency of the United States Government or under a contract with an agency of the United States Government.

[x] No [ ] Yes, the name of the U.S. Government agency and the Government contract number are: ______

Respectfully submitted,

Date March 4, 2004

SIGNATURE Ign R de la Cr
TYPED OR PRINTED NAME Ignacio Perez de la Cruz
TELEPHONE (212) 497-3613

REGISTRATION NO. (*if appropriate*) 55,535
Docket Number: WIBL-P60-035

USE ONLY FOR FILING A PROVISIONAL APPLICATION FOR PATENT

I hereby certify that this correspondence is being deposited with the U.S. Postal Service as Express Mail, Airbill No. EJ624594035US, in an envelope addressed to: MS Provisional Patent Application, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, on the date shown below.

Dated: March 4, 2004 Signature: Ign R de la Cr (Ignacio Perez de la Cruz)

9386980_1.DOC

PTO/SB/17 (10-03)
Approved for use through 7/31/2006. OMB 0651-0032
U.S. Patent and Trademark Office; U.S. DEPARTMENT OF COMMERCE

FEE TRANSMITTAL for FY 2004

Effective 10/01/2003, Patent fees are subject to annual revision.

☐ Applicant claims small entity status. See 37 CFR 1.27

TOTAL AMOUNT OF PAYMENT ($) 160.00

Complete if Known	
Application Number	Not Yet Assigned
Filing Date	March 4, 2004
First Named Inventor	Christopher T. Harbison
Examiner Name	Not Yet Assigned
Art Unit	N/A
Attorney Docket No.	WIBL-P60-035

METHOD OF PAYMENT (check all that apply)

☐ Check ☐ Credit Card ☐ Money Order ☐ Other ☐ None

[X] Deposit Account:

Deposit Account Number: 18-1945

Deposit Account Name: Ropes & Gray LLP

The Director is authorized to: (check all that apply)

[X] Charge fee(s) indicated below [X] Credit any overpayments

[X] Charge any additional fee(s) or any underpayment of fee(s)

☐ Charge fee(s) indicated below, **except for the filing fee**

to the above-identified deposit account.

FEE CALCULATION

1. BASIC FILING FEE

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
1001	770	2001	385	Utility filing fee	
1002	340	2002	170	Design filing fee	
1003	530	2003	265	Plant filing fee	
1004	770	2004	385	Reissue filing fee	
1005	160	2005	80	Provisional filing fee	160.00

SUBTOTAL (1) ($) 160.00

2. EXTRA CLAIM FEES FOR UTILITY AND REISSUE

			Extra Claims		Fee from below		Fee Paid
Total Claims		-** =		x		=	
Independent Claims		-** =		x		=	
Multiple Dependent						=	

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description
1202	18	2202	9	Claims in excess of 20
1201	86	2201	43	Independent claims in excess of 3
1203	290	2203	145	Multiple dependent claim, if not paid
1204	86	2204	43	** Reissue independent claims over original patent
1205	18	2205	9	** Reissue claims in excess of 20 and over original patent

SUBTOTAL (2) ($) 0.00

***or number previously paid, if greater; For Reissues, see above*

FEE CALCULATION (continued)

3. ADDITIONAL FEES

Large Entity Fee Code	Large Entity Fee ($)	Small Entity Fee Code	Small Entity Fee ($)	Fee Description	Fee Paid
1051	130	2051	65	Surcharge – late filing fee or oath	
1052	50	2052	25	Surcharge – late provisional filing fee or cover sheet.	
1053	130	1053	130	Non-English specification	
1812	2,520	1812	2,520	For filing a request for *ex parte* reexamination	
1804	920*	1804	920*	Requesting publication of SIR prior to Examiner action	
1805	1,840*	1805	1,840*	Requesting publication of SIR after Examiner action	
1251	110	2251	55	Extension for reply within first month	
1252	420	2252	210	Extension for reply within second month	
1253	950	2253	475	Extension for reply within third month	
1254	1,480	2254	740	Extension for reply within fourth month	
1255	2,010	2255	1,005	Extension for reply within fifth month	
1401	330	2401	165	Notice of Appeal	
1402	330	2402	165	Filing a brief in support of an appeal	
1403	290	2403	145	Request for oral hearing	
1451	1,510	1451	1,510	Petition to institute a public use proceeding	
1452	110	2452	55	Petition to revive – unavoidable	
1453	1,330	2453	665	Petition to revive - unintentional	
1501	1,330	2501	665	Utility issue fee (or reissue)	
1502	480	2502	240	Design issue fee	
1503	640	2503	320	Plant issue fee	
1460	130	1460	130	Petitions to the Commissioner	
1807	50	1807	50	Processing fee under 37 CFR 1.17(q)	
1806	180	1806	180	Submission of Information Disclosure Stmt	
8021	40	8021	40	Recording each patent assignment per property (times number of properties)	
1809	770	2809	385	Filing a submission after final rejection (37 CFR 1.129(a))	
1810	770	2810	385	For each additional invention to be examined (37CFR 1.129(b))	
1801	770	2801	385	Request for Continued Examination (RCE)	
1802	900	1802	900	Request for expedited examination of a design application	

Other fee (specify) ______

*Reduced by Basic Filing Fee Paid **SUBTOTAL (3)** ($) 0.00

SUBMITTED BY				(Complete (*if applicable*))	
Name (Print/Type)	Ignacio Perez de la Cruz	*Registration No. (Attorney/Agent)*	55,535	*Telephone*	(212) 497-3613
Signature	Ig Pr de la Cr			*Date*	March 4, 2004

I hereby certify that this correspondence is being deposited with the U.S. Postal Service as Express Mail, Airbill No. EJ624594035US, in an envelope addressed to: MS Provisional Patent Application, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, on the date shown below.

Dated: March 4, 2004 Signature: Ig Pr de la Cr (Ignacio Perez de la Cruz)

9387016_1.DOC

Application No. (if known): Not Yet Assigned   Attorney Docket No.: WIBL-P60-035

Certificate of Express Mailing Under 37 CFR 1.10

I hereby certify that this correspondence is being deposited with the United States Postal Service as Express Mail, Airbill No. EJ624594035US in an envelope addressed to:

MS Provisional Patent Application
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

on March 4, 2004
Date

Ignacio Perez de la Cruz
Signature

Ignacio Perez de la Cruz
Typed or printed name of person signing Certificate

Note: Each paper must have its own certificate of mailing, or this certificate must identify each submitted paper.

Application Data Sheet (3 pages)
Application (37 pages)
Drawings (38 sheets)

Transcriptional Regulatory Codes of Eukaryotic Genomes and Methods Thereof

Christopher T. Harbison[1,2*], D. Benjamin Gordon[1*], Tong Ihn Lee[1], Nicola J. Rinaldi[1,2], Kenzie Macisaac[3], Timothy Danford[3], Nancy M. Hannett[1], Jean-Bosco Tagne[1], David B. Reynolds[1], Jane Yoo[1], Ezra G. Jennings[1], Julia Zeitlinger[1], Manolis Kellis[1], Alex Rolfe[3],Ken T. Takusagawa[3], David K. Gifford[3], Ernest Fraenkel[1,3†] and Richard A. Young[1,2†]

[*]These authors contributed equally to this work

[1]*Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA*

[2]*Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

[3]*MIT Computer Science and Artificial Intelligence Laboratory, 200 Technology Square, Cambridge, MA 02139, USA*

†To whom correspondence should be addressed:

RAY: E-mail young@wi.mit.edu, EF: E-mail efraenkel@wi.mit.edu

Supporting Online Material

Address: http://web.wi.mit.edu/young/regulatory_code

Username: reviewer

Password: code

We report here a first draft of a genome's transcriptional regulatory code, the set of sequences utilized by DNA-binding regulators to control genome expression programs. The code was derived by combining data on genomic binding locations for transcriptional regulators in yeast cells grown in multiple environmental conditions, knowledge of genome sequence conservation and prior evidence for regulator-DNA interactions. We discuss new insights into global transcriptional regulation that are revealed by the code, including the organization of regulatory elements in promoters and the environment-dependent use of these elements by regulators. We find that environment-specific use of the regulatory code predicts mechanistic models for the function of a large population of yeast's transcriptional regulators.

Genome sequences contain information necessary to control gene expression programs and specify protein and other gene products. DNA-binding transcriptional regulators interpret the genome's regulatory code by binding to specific sequences to induce or repress gene expression [1-3]. Substantial portions of genome sequence are believed to be regulatory [4-8], but the DNA sequences that actually contribute to the regulatory code are ill-defined. In contrast, the triplet code used to translate nucleotide sequences into protein molecules is well known [9-11]. Knowledge of the genome's transcriptional regulatory code could provide new insights into the principles that govern global gene regulation.

Comparative genomics has recently been used to identify functional sequence elements in the yeast genome [4,5,12-14]. Comparative analysis of the genome sequences of multiple yeast species revealed phylogenetically conserved sequences, and these sequences were used to facilitate identification of genes and putative regulatory elements. Conserved sequence information alone does not reveal, however, the subset of sequences that are bound by transcriptional regulators, the identity of the binding regulators, or the conditions under which the regulators occupy their binding sites.

The set of DNA sequences that play key roles in transcriptional regulation can be deduced by identifying the genomic binding locations for individual regulatory proteins [15-20] and by searching for DNA sequences shared by the bound sites using statistical algorithms [21-23]. However, genomic binding information is available only for a subset of regulators in any organism. The genomic binding sites of 106 transcriptional regulators have been identified in yeast grown in rich medium [16], but there are approximately 200 transcriptional regulators encoded in the yeast genome, and many of these are known to function under conditions other than a rich medium environment.

Proteome-Genome Interactions

To elucidate a substantial portion of the yeast genome's transcriptional regulatory code, we used genome-wide location analysis [15] to determine the genomic occupancy of 203 DNA-binding transcriptional regulators in rich media conditions and, for 85 of these regulators, in at least one of twelve other environmental conditions. The 203 transcriptional regulators were identified by searching the YPD and MIPS databases [24-26] for known and predicted transcriptional regulators and nucleic acid binding proteins. These are likely to include nearly all of the DNA-binding transcriptional regulators encoded in the yeast genome. We selected regulators for profiling in a specific environment if they were essential for growth in that environment or if there was other evidence implicating them in regulation of gene expression in that environment (see online supporting data). The complete set of transcriptional regulators and the conditions under which they were profiled is listed in Supplementary Table S1, and all genome-wide location datasets are available at http://web.wi.mit.edu/young/regulatory_code.

The genome-wide location data identified 11,000 unique interactions between regulators and promoter regions. There was a broad distribution in the number of promoter regions bound by the transcriptional regulators (Figure 1a); the average regulator bound to approximately 55 promoter regions ($P < 0.001$). Among the group of regulators that bound the most promoters, a disproportionate number were nuclear regulators known to be highly abundant [27], including Abf1, Cbf1, Reb1 and Ste12. Twenty putative regulators bound no promoter regions at high confidence ($P < 0.001$), suggesting that they function under conditions that were not explored (see below) or that they are not genuine transcriptional regulators. The location data also reveal that nearly one-half of the genes that are bound by regulators were bound by two or more

regulators (Supplementary Fig. S2), and that a subset of genes are bound by a very large number of regulators, suggesting that they are regulated by multiple signals and subjected to combinatorial control.

For the 85 regulators profiled in rich media and at least one other environment, we found that a substantial number of promoter regions that were not bound in rich media were bound in the other environments. Approximately one thousand regulator-promoter region interactions identified in the other environments involved promoters that were not bound by any regulator under rich growth conditions. For example, a comparison of the promoter regions bound by 34 regulators in cells grown under rich or amino acid starved environments (Figure 1b) reveals that there are substantial differences in the number of promoters bound by many of these regulators. These results underscore the importance of the role of environmental conditions in governing DNA-regulator interactions, and we explore environment-dependent global gene regulation in more detail below.

DNA Binding Site Sequences

We combined genome-wide location data, phylogenetically conserved sequence information, and prior knowledge of regulator binding sequences to create a database of DNA binding sequences for transcriptional regulators (Figure 2a). To accomplish this, we first combined genome-wide location and sequence conservation data to predict binding site specificities. For regulators where we failed to predict binding site specificities at high confidence, we used evidence for binding site sequences obtained from the literature, where available.

Genome-wide location analysis identifies regions of the genome that are physically occupied by specific regulators in vivo, but does not identify the precise DNA sequences that serve as recognition sites. Numerous algorithms have been

developed to discover binding sites in a set of identified sequences. We designed a strategy to improve the yield of the discovery process by combining the results from six programs, three of which incorporate information from phylogenetic conservation [21-23,28] ; for a detailed description of the method, see supplementary data. Using this approach, we identified binding site sequence specificities for the 151 regulators that bound more than ten promoter regions. Of these, 68 met a high threshold criteria for significance (see Methods and Supplementary Table S2).

We compared the discovered binding site sequence specificities with previous knowledge by generating a list of binding sequences and sequence motifs for each regulator that was derived solely from the evidence for protein-DNA interactions as contained in the databases TRANSFAC, SCPD and YPD [24,25,29,30] (Supplementary Table S2). For 39 transcriptional regulators, we found that we identified a DNA binding site sequence that was identical to, or shared significant homology with, previously described binding sequences. For example, the DNA binding sequence motifs for Abf1, Gal4, Gcn4, Leu3 and Ste12 were rediscovered (Figure 2b). For 21 additional regulators, the DNA binding site sequence we predict was completely novel; the newly discovered sequence had not been described previously. Aft2, Rds1, Snt2, Stb4 and YDR026C were among the regulators for which novel binding sequences were discovered (Figure 2b). For 8 regulators, the DNA binding sequence we predict differs from the sequence described in the literature. This discrepancy may be due to different growth environments used in the different studies, to noise in the genome-wide location data, or to limitations in the analytical methods used here or in previous studies.

The literature suggests binding site sequences for 15 transcriptional regulators for which we did not predict a binding site sequence. We added this literature evidence to our database of DNA binding sequences for transcriptional regulators (Supplementary Table S2) and used it in subsequent analyses.

We note that the regulators bind to only a fraction of the sites in the genomic DNA that contain the predicted recognition sequence. To test whether this observation is due to the limitations of our methods or whether it reflects a fundamental aspect of genomic regulation, we investigated whether bound sites were under more selective pressure than sites that were not bound. We used DNA binding site specificities listed in the TRANSFAC database in this analysis to eliminate the sequence conservation bias inherent in the binding sites specificities determined as described above. Figure 2c shows that the genomic loci that both match specificities in the TRANSFAC database and were bound in our assay are more conserved in other *sensu stricto Saccharomyces* species than loci that match the TRANSFAC specificities but were not bound. For example, the regulator Hsf1 binds to only 23/63 genomic loci that match the TRANSFAC specificity. While 78% of the bound sequences are conserved, only 13% of the remaining sites are conserved.

There are both positive and negative regulatory mechanisms that are likely to account for the observation that DNA binding regulators bind only a subset of the sites in genomic DNA that contain that regulator's recognition sequence. Cis-acting regulatory DNA sequences frequently contain binding sites for multiple regulators that stimulate cooperative binding through protein-protein interactions or that alter local DNA structure. The presence of multiple regulators of this type might be expected to be conserved in related species, accounting for the relative selective pressure observed for the bound sequences. It is also the case that the binding of a protein to certain sites in the genome can be occluded by the presence of another protein. The proteins associated with genomic DNA include transcriptional regulators, the transcription apparatus, the DNA replication apparatus, histones and other chromatin-associated proteins, and chromatin modifying complexes. These are all likely to affect the relative occupancy of specific DNA sequences by transcriptional regulators.

We also note that some DNA binding regulators can occupy DNA sequences in the absence of a discernable binding sequence for that regulator. There are several examples of proteins that occupy specific sites in the absence of an appropriate sequence. When Tec1 is present in cells, Ste12 can be found associated with DNA with Tec1 binding sequences [31]. The Hir1/Hir2 corepressor complex, which is recruited to the promoters of histone genes, does not bind to a sequence that is conserved in each of the histone promoters [32]. The Origin of DNA Replication Complex (ORC) occupies specific genomic sites [33], but a consensus sequence that would meet the confidence criteria used here has not been found. These latter two examples demonstrate that more than one unique DNA sequence can provide a functional binding site for transcriptional and DNA synthesis regulators, and emphasize the importance of in vivo binding data in discovering functional elements in the genome.

Transcriptional regulatory code

Gene expression programs depend on recognition of specific DNA sequences by transcriptional regulatory proteins, so the set of DNA sequences that are bound by these proteins should reveal the genome's transcriptional regulatory code. We have constructed a draft of the yeast genome's transcriptional regulatory code by identifying the positions of the sequences that are bound by regulators in vivo in the genome sequence. To populate this draft of the regulatory code with high confidence information, we further restricted the map to include sequences that were not only bound, but were also conserved in at least two other *sensu stricto Saccharomyces* genomes [4]. Portions of the draft code is displayed in Figure 3, and the complete map can be found at the authors' website. Because the information used to construct the map includes binding data from many different growth environments, the code describes transcriptional regulatory potential within the genome.

The first level of information contained within the regulatory code is simply the association of genes with the regulators responsible for their transcriptional control. This level of information reveals insights into the control of underlying biological processes. For example, find that the promoter of *BAP2*, which encodes an extracellular amino acid transporter, is bound by the amino acid biosynthetic regulators Gcn4 and Leu3. Similarly, we identify binding of the regulator of respiration, Hap5, to a site upstream of *COX4*, a component of the respiratory electron transport chain. The ribosomal regulator Rap1 is bound to the small ribosomal subunit component gene RPS18A. Knowledge of which regulators control which genes is fundamental in understanding particular aspects of biology. However, in many cases, the identity of regulators and the arrangement of their binding sites within promoters suggests additional levels of organization (discussed below).

The distribution of binding sites for transcriptional regulators reveals there are constraints on the organization of promoters in the yeast genome (Supplemental Fig. S1). Binding sites are not uniformly distributed over the promoter regions, but rather show a sharply peaked distribution. Very few sites are located in the region 80 bp upstream of protein coding sequences. This region typically includes the transcription start site and is bound by the transcription initiation apparatus; RNA polymerase alone has been shown to occupy approximately 60 bp. The vast majority (73%) of the transcriptional regulator binding sites lie between 100 and 500 bps upstream of the protein coding sequence. It appears that yeast transcriptional regulators function better at short distances along the linear DNA, a property that reduces the potential for inappropriate activation of nearby genes and allows evolution to select for organisms that dispense with lengthy intergenic regions.

Promoter Architectures

We note that there are specific arrangements of DNA binding site sequences within promoters, and that these promoter architectures can provide clues to regulatory mechanisms. We identify four such distinctive arrangements (Figure 4) and discuss their biological implications below.

Single regulator architecture. The presence of a DNA binding site for a single regulator is the simplest promoter architecture (Figure 4). Seven hundred sixty-three promoter regions were found to have single regulator promoter architecture (see online supporting data). These promoter regions include the binding sites for 75 of the 203 regulators. We expect that sets of genes containing a binding site for a single regulator may be involved in a common function, and this is the case for genes bound by 25 regulators (Supplementary Table S3). Analysis of expression data confirmed that for 15 of these regulators, there is also strong correlation between single regulator promoters and changes in gene expression (Supplementary Table S3). We do not expect correlation in all cases since expression is known to be regulated at levels other than transcription initiation. For example, mechanisms are known that control both mRNA stability and export from the nucleus [34].

Multiple regulator architecture. Promoters with multiple regulator architecture contain binding sites for two or more different regulators (Figure 4). We find that 533 promoter regions contain sequences corresponding to two or more binding site motifs. These regions are bound by 99 different regulators. This promoter arrangement implies that the gene may be subject to combinatorial regulation, and we expect that in many cases the various regulators can be used to execute differential responses to different growth conditions. Indeed, we note that many of the genes in this category encode products that are required for multiple metabolic pathways and are regulated in an environment-specific fashion. For example, the promoter of *IDP1* gene, which is involved in energy production through oxidation, is bound not only by Hap4 and Rtg3,

regulators of carbon metabolism, but also by Gln3 and Gcn4. These latter proteins regulate amino acid biosynthetic pathways that require the same simple carbon compounds generated by genes under the control of Idp1.

Repetitive motif architecture. While some promoter regions contain only a single copy of a particular binding site sequence, others contain repeats of these sites (Figure 4). Multiple binding sites have been shown to be necessary for stable binding by the regulator Dal80 [35], and given the small size of typical binding site sequences, this requirement is likely for additional regulators. This repetitive promoter architecture can also allow for a graded transcriptional response, as has been observed for the *HIS4* gene [36,37]. The presence of repeated binding site sequences for a single regulator was observed for 71 regulators across 408 promoter regions (Figure 4). We note that a number of regulators, including Dig1, Mbp1, and Swi6 show a statistically significant preference for repetitive motifs (Supplementary Table S4).

Co-occurring regulator architecture. We searched the set of multiple regulator promoters to identify pairs of transcription factors whose binding sites occur more frequently within the same promoter regions than would be expected by chance (Table S5). There are ninety-three such co-occurring pairs of regulators ($P < 0.005$). We expect three types of relationships among such regulators. The first type is exemplified by the regulatory pair Gcn4 and Leu3. In this case, the pair of regulators share a functional overlap in their regulatory roles. Specifically, Gcn4 is a general regulator of amino acid biosynthetic genes whereas Leu3 regulates a smaller set of genes involved solely in the metabolism of aliphatic amino acids. This arrangement apparently allows cells the option of co-ordinately regulating either the complete set of amino acid biosynthetic genes or a specific subset. A second type of co-occurring regulator pairs consists of regulators that form physical associations. A well studied example of such a pair is Swi6 and Mbf1, which together form the MBF heterodimeric cell cycle

regulatory complex. Finally, where pairs of regulators do not interact physically, but bind to the same DNA sequence, the binding site is polysemic, conveying more than one meaning. In many cases, such regulator pairs share homologous DNA-binding domains. The regulators Fkh1 and Fkh2, for example, are known to have overlapping but distinct functions, and this is reflected in their binding profiles.

Environment-specific Binding of Regulatory Sequences

By conducting genome-wide binding experiments for some regulators under multiple cell growth conditions, we learned that regulator binding to a subset of the regulatory sequences is highly dependent on the environmental conditions of the cell. We observed four common patterns of regulator binding behaviour (Figure 5, Supplementary Table S6). Prior information about the regulatory mechanisms employed by well-studied regulators in each of the four groups suggests how to account for the environment-dependent binding behaviour of the other regulators.

"Condition invariant" regulators bind essentially the same set of promoters (within the limitations of noise) in two different growth environments (Figure 5). Leu3, which is known to regulate genes involved in amino acid biosynthesis, is among the best studied of the regulators in this group. Activation of Leu3-regulated genes has been shown to be independent of Leu3 binding, but requires association of a leucine metabolic precursor to convert it from negative to positive regulator [38-40]. We note that other zinc cluster type regulators that show "condition invariant" behaviour are known to be regulated in a similar manner. Thus, it is reasonable to propose that the activation or repression functions of some of the other regulators in this class will be independent of DNA binding.

"Condition enabled" regulators do not bind the genome detectably under one condition, but bind a substantial number of promoters with a change in environment.

Msn2 is among the best-studied regulators in this class, and the mechanisms involved in Msn2-dependent transcription provide clues to how the other regulators in that class may operate. Msn2 is known to be excluded from the nucleus when cells grow in the absence of stresses, but accumulates rapidly in the nucleus when cells are subjected to stress [41-43]. This condition-enabled behaviour was also observed for the thiamine biosynthetic regulator Thi2, the nitrogen regulator Gat1 and the developmental regulator Rim101. We postulate that each of these transcriptional regulators is regulated by nuclear exclusion or by another mechanism that would cause this extreme version of condition-specific binding.

"Condition expanded" regulators bind to a core set of target promoters under one condition, but bind an expanded set of promoters under another condition. Gcn4 is the best-studied of the regulators that fall into this "expanded" class. The levels of Gcn4 are reported to increase 6-fold when yeast are introduced into media with limiting nutrients [44], due largely to increased nuclear protein stability [42,45,46], and under this condition we find Gcn4 binds to an expanded set of genes. Interestingly, the probes bound when Gcn4 levels are low contain better matches to the known Gcn4 binding site than probes that are bound exclusively at higher protein concentrations, consistent with a simple model for specificity based on intrinsic protein affinity and protein concentration (Supplementary Figure S1). The expansion of binding sites by many of the regulators in this class likely reflects increased levels of the regulator available for DNA binding.

"Condition altered" regulators exhibit altered preference for the set of promoters bound in two different conditions. Ste12 is the best studied of the regulators whose binding behaviour falls into this "altered" class. The specificity of Ste12 is thought to depend, in part, on its interactions with other regulators whose availability is environment-dependent. For example, the binding site preference of Ste12 changes when Tec1 is present, apparently because Tec1 interacts with Ste12 and has its own

DNA-binding site specifity [31,47]. This condition-altered behaviour was also observed for the transcriptional regulators Aft2, Pho4, and Ume6. We postulate that the binding specificity of many of the transcriptional regulators may be altered through interactions with other regulators or through modifications (e.g., chemical) that are environment-dependent.

We note that classification of regulator behaviour is dependent on the environmental conditions selected for comparison, and that some regulators fall into multiple categories. This is consistent with the understanding that multiple types of regulatory mechanisms are often associated with regulators, and differentially modulate regulator behaviour. We anticipate that future experiments in additional environments will provide a more comprehensive understanding of regulator dynamics.

Challenges for Future Drafts of the Regulatory Code

We have used extensive *in vivo* binding data, conserved sequence information and prior knowledge of regulator-DNA interactions to construct a first draft of the transcriptional regulatory code of a eukaryote cell. We anticipate that future revisions will be facilitated by collecting more experimental data, by testing models that emerge from the data, and by developing improved computational algorithms to integrate various data types. It will be valuable to collect genome-wide binding data for DNA-binding regulators and chromatin regulators in cells grown under additional environments. Experimental tests of models for regulator functions predicted by their environment-dependent binding behaviour will provide new insights into the regulatory mechanisms involved in control of global gene expression programs. Knowledge of the environment-dependent changes in the abundance, modification state and intracellular compartmentalization of transcriptional regulators will also be valuable, although collecting this evidence will be challenging because transcriptional regulators are

among the least abundant of cellular proteins. Frequent sampling of genome-wide expression data obtained as cells are placed in the new environments will permit investigators to integrate binding and expression data to explain how dynamic changes in gene expression programs are regulated. New computational algorithms will play a key role in improved binding site sequence prediction and allowing integration of various data types to reveal how the transcriptional regulatory code controls gene expression programs under diverse conditions.

Methods

Strain Information

Strains were created for each of the 204 regulators in which a repeated Myc epitope coding sequence was integrated into the endogenous gene encoding the regulator. PCR constructs containing the Myc epitope coding sequence and a selectable marker flanked by regions of homology to either the 5' or 3' end of the targeted gene were transformed into the W303 yeast strain Z1256. Genomic integration and expression of the epitope-tagged protein were confirmed by PCR and Western blotting, respectively.

Genome wide location analysis

Genome-wide location analysis was performed as previously described. Bound proteins were formaldehyde-crosslinked to DNA in vivo, followed by cell lysis and sonication to shear DNA. Crosslinked material was immunoprecipitated with an anti-myc antibody, followed by reversal of the crosslinks to separate DNA from protein. Immunoprecipitated DNA and DNA from an unenriched sample were amplified and differentially fluorescently labeled by ligation-mediated PCR. These samples were

hybridized to a microarrray consisting of spotted PCR products representing the intergenic regions of the *S. cerevisiae* genome. Relative intensities of spots were used as the basis for an error model that assigns a probability score (*P* value) to binding interactions.

Growth environments

We profiled all 203 regulators in rich medium. In addition, we profiled 85 regulators in at least one other environmental condition. The list of regulators is given in Supplementary Table S1 and more information about the selection of these regulators and the environmental conditions used can be found at the author's web site.

Motif discovery

We used six methods to identify the specific sequences bound by regulators: AlignACE, MEME, MDscan, the method of Kellis et al. and two additional new methods that incorporate conservation data: MEME_c and CONVERGE. MEME_c uses the existing MEME program without change, but applies it to a modified set of sequences in which non-conserved bases were replaced with the letter "N". CONVERGE, is a novel EM-based algorithm that takes a set of multiple sequence alignments (MSA) as input instead of a set of sequences. Each MSA contains the available conservation information for a single probe across the *sensu stricto* species. Rather than searching for sites that are identical across multiple species, as is the case for MEME_c, CONVERGE searches for loci where all aligned sequences are consistent with the same specificity model. CONVERGE is described in greater detail in the supplemental material and will be published elsewhere.

We used XXX statistics to judge the significance of motifs A, B, C ... (described in the supplementary material). To determine the appropriate thresholds for these measures, we applied each program to sets of randomly selected probes and calculated the empirical probability distribution for each program to find a motif with the given score. A motif was accepted if any one of that scores had a p-value ≤ 0.001 when compared distribution of the same score observed in 50 randomization runs with the same program. Significant motifs for the same factor were clustered and averaged.

Regulatory code

Potential binding sites were included in the map of the regulatory code if they satisfied two criteria. First, a locus had to match the specificity model for a regulator in the *Saccharomyces cerevisiae* genome and at least two other *sensu stricto cerevisiae* genomes with a score $\geq$70% of the maximum possible. Second, the locus had to lie in an intergenic region that also contained a probe bound by the corresponding factor ($p \leq 0.001$).

Figure 1. Genome-wide Distribution of Transcriptional Regulators. Genome-wide location analysis was used to determine the genomic occupancy of 203 DNA-binding transcriptional regulators in rich media conditions and, for 85 of these regulators, in at least one of twelve other environmental conditions. Additional information on experimental protocols, the regulators under study, raw location data, and data analysis methods are available in supplementary data and on the author's website[48]. The data selected for further analysis met a 0.001 *P* value threshold (Lee et al., 2002). a. Distribution of the number of promoter regions bound per regulator. For regulators profiled under multiple conditions, the union of promoter regions bound under all conditions is reported (blue). An average of randomized distributions for the same set of *P* values randomly assigned among regulators and promoter regions is shown in pink. b. Pairwise comparison of the number of promoter regions bound under two different conditions for 25 regulators. Dark blue bars represent the number of promoter regions bound under growth in rich medium; light blue bars represent the number of promoter regions bound under growth in amino acid starvation medium.

Figure 2. Binding Site Sequences for Yeast Transcriptional Regulators. a. Experimental procedure. Data from location analysis experiments were subjected to computational analyses to determine regulator binding site sequence specificities. These sequences were compared to and supplemented with published sequence specificities. Phylogenetic conservation information was used both to determine, in part, the binding sequences and in selecting the set of final sequences for inclusion in the map of the regulatory code. b. Specificity models that were "rediscovered" using our location data (left) as well as for newly discovered sequences (right). The height of the letter in each logo

represents the frequency of the given nucleotide at that position within the collection of binding specificities. c. Conservation of bound motifs. The degreee of conservation of genomic sites that match a TRANSFACE matrix and were bound (red) or unbound (dark blue). The distribution of conservation values for all possible six-base pair sequences is shown for comparison (cyan).

Figure 3. Yeast Transcriptional Regulatory Code. A genomic map of regulatory code sequences. Regions of Chromosomes II, IV and VII are shown including the locations of conserved DNA sequences bound *in vivo* by transcriptional regulators. Genes are shown as grey rectangles with arrows indicating direction of transcription; binding sequences are shown as coloured boxes.

Figure 4. Yeast promoter architectures. Four classes of arrangements present in promoter regions are depicted.

Figure 5. Environment-specific Utilization of Transcriptional Regulatory Code. Four patterns of genome-wide binding behaviour are depicted in a graphic representation on the left, where transcriptional regulators are represented by a coloured circle, a representative set of gene/promoters is represented above and below the regulator, and lines between the regulator and the gene/promoters represent binding events. Specific examples of the environment-dependent behaviours are depicted in the middle. Coloured circles represent regulators and coloured boxes represent DNA binding sequences present within promoter regions.

Figures 8-39. Specific Embodiments of the methods described herein.

1. Jacob, F. & Monod, J. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* **3**, 318-56 (1961).
2. Gilbert, W. & Muller-Hill, B. The lac operator is DNA. *Proc Natl Acad Sci U S A* **58**, 2415-21 (1967).
3. Ptashne, M. Specific binding of the lambda phage repressor to lambda DNA. *Nature* **214**, 232-4 (1967).
4. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241-54 (2003).
5. Cliften, P. et al. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science* **301**, 71-6 (2003).
6. Venter, J. C. et al. The sequence of the human genome. *Science* **291**, 1304-51 (2001).
7. Lander, E. S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
8. Waterston, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).
9. Matthaei, J. H., Jones, O. W., Martin, R. G. & Nirenberg, M. W. Characteristics and composition of RNA coding units. *Proc Natl Acad Sci U S A* **48**, 666-77 (1962).
10. Crick, F. H., Barnett, L., Brenner, S. & Watts-Tobin, R. J. General nature of the genetic code for proteins. *Nature* **192**, 1227-32 (1961).
11. Khorana, H. G. Polynucleotide synthesis and the genetic code. *Fed Proc* **24**, 1473-87 (1965).
12. Pritsker, M., Liu, Y. C., Beer, M. A. & Tavazoie, S. Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res* **14**, 99-108 (2004).
13. Wang, T. & Stormo, G. D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19**, 2369-80 (2003).
14. Blanchette, M. & Tompa, M. FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res* **31**, 3840-2 (2003).
15. Ren, B. et al. Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306-9. (2000).
16. Lee, T. I. et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804. (2002).
17. Iyer, V. R. et al. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* **409**, 533-8. (2001).
18. Simon, I. et al. Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell* **106**, 697-708. (2001).
19. Horak, C. E. et al. Complex transcriptional circuitry at the G1/S transition in Saccharomyces cerevisiae. *Genes Dev* **16**, 3017-33 (2002).
20. Lieb, J. D., Liu, X., Botstein, D. & Brown, P. O. Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. *Nat Genet* **28**, 327-34 (2001).
21. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* **16**, 939-45. (1998).

22. Bailey, T. L. & Elkan, C. The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol* **3**, 21-9 (1995).
23. Liu, X. S., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol* **20**, 835-9 (2002).
24. Costanzo, M. C. et al. YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res* **29**, 75-9 (2001).
25. Hodges, P. E., McKee, A. H., Davis, B. P., Payne, W. E. & Garrels, J. I. The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res* **27**, 69-73 (1999).
26. Mewes, H. W., Albermann, K., Heumann, K., Liebl, S. & Pfeiffer, F. MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res* **25**, 28-30 (1997).
27. Ghaemmaghami, S. et al. Global analysis of protein expression in yeast. *Nature* **425**, 737-41 (2003).
28. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**, 1205-14 (2000).
29. Knuppel, R., Dietze, P., Lehnberg, W., Frech, K. & Wingender, E. TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. *J Comput Biol* **1**, 191-8 (1994).
30. Zhu, J. & Zhang, M. Q. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* **15**, 607-11 (1999).
31. Zeitlinger, J. et al. Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling. *Cell* **113**, 395-404 (2003).
32. Spector, M. S., Raff, A., DeSilva, H., Lee, K. & Osley, M. A. Hir1p and Hir2p function as transcriptional corepressors to regulate histone gene transcription in the Saccharomyces cerevisiae cell cycle. *Mol Cell Biol* **17**, 545-52 (1997).
33. Wyrick, J. J. et al. Genome-wide distribution of ORC and MCM proteins in S. cerevisiae: high-resolution mapping of replication origins. *Science* **294**, 2357-60 (2001).
34. Hieronymus, H. & Silver, P. A. Genome-wide analysis of RNA-protein interactions illustrates specificity of the mRNA export machinery. *Nat Genet* **33**, 155-61 (2003).
35. Cunningham, T. S. & Cooper, T. G. The Saccharomyces cerevisiae DAL80 repressor protein binds to multiple copies of GATAA-containing sequences (URSGATA). *J Bacteriol* **175**, 5851-61 (1993).
36. Donahue, T. F., Daves, R. S., Lucchini, G. & Fink, G. R. A short nucleotide sequence required for regulation of HIS4 by the general control system of yeast. *Cell* **32**, 89-98 (1983).
37. Arndt, K. & Fink, G. R. GCN4 protein, a positive transcription factor in yeast, binds general control promoters at all 5' TGACTC 3' sequences. *Proc Natl Acad Sci U S A* **83**, 8516-20 (1986).
38. Brisco, P. R. & Kohlhaw, G. B. Regulation of yeast LEU2. Total deletion of regulatory gene LEU3 unmasks GCN4-dependent basal level expression of LEU2. *J Biol Chem* **265**, 11667-75 (1990).

39. Friden, P., Reynolds, C. & Schimmel, P. A large internal deletion converts yeast LEU3 to a constitutive transcriptional activator. *Mol Cell Biol* **9**, 4056-60 (1989).
40. Kirkpatrick, C. R. & Schimmel, P. Detection of leucine-independent DNA site occupancy of the yeast Leu3p transcriptional activator in vivo. *Mol Cell Biol* **15**, 4021-30 (1995).
41. Beck, T. & Hall, M. N. The TOR signalling pathway controls nuclear localization of nutrient-regulated transcription factors. *Nature* **402**, 689-92 (1999).
42. Chi, Y. et al. Negative regulation of Gcn4 and Msn2 transcription factors by Srb10 cyclin-dependent kinase. *Genes Dev* **15**, 1078-92. (2001).
43. Gorner, W. et al. Nuclear localization of the C2H2 zinc finger protein Msn2p is regulated by stress and protein kinase A activity. *Genes Dev* **12**, 586-97 (1998).
44. Albrecht, G., Mosch, H. U., Hoffmann, B., Reusser, U. & Braus, G. H. Monitoring the Gcn4 protein-mediated response in the yeast Saccharomyces cerevisiae. *J Biol Chem* **273**, 12696-702. (1998).
45. Meimoun, A. et al. Degradation of the transcription factor Gcn4 requires the kinase Pho85 and the SCF(CDC4) ubiquitin-ligase complex. *Mol Biol Cell* **11**, 915-27. (2000).
46. Kornitzer, D., Raboy, B., Kulka, R. G. & Fink, G. R. Regulated degradation of the transcription factor Gcn4. *Embo J* **13**, 6021-30. (1994).
47. Baur, M., Esch, R. K. & Errede, B. Cooperative binding interactions required for function of the Ty1 sterile responsive element. *Mol Cell Biol* **17**, 4330-7 (1997).
48. Tessier, D. et al. *A DNA Microarrays Fabrication Strategy for Research Laboratories.* (eds. Rehm, H. & Reed, G.) (Wiley-VCH, Weinheim, Germany, 2002).

Description of Supplementary Tables and Figures:

Supplementary Table S1. List of all factors profiled and conditions used.

Supplementary Table S2. List of all binding sequences used for creating map.

Supplementary Table S3. Table of Single Regulator MIPs annotations.

Supplementary Table S4. List of regulators and genes with repetitive motif architecture.

Supplementary Table S5. List of regulators and genes with co-occurring motif architecture.

Supplementary Table S6. Classification of all regulators into binding behaviour categories.

Supplementary Figure S1. Distance to ATG.

Supplementary Figure S2. Gcn4 motifs.

Supplementary Table S1

A1	Kre33	Smp1
Abf1	Kss1 [5,6]	Snf1
Abt1	Leu3 [3]	Snt2
Aca1	Mac1 [1]	Sok2 [5]
Ace2	Mal13	Spt10
Adr1 [3,7]	Mal33 [1,2]	Spt2
Aft2 [1,2]	Mbf1	Spt23
Arg80 [3]	Mbp1 [1,2]	Srd1
Arg81 [3]	Mcm1 [5,6]	Stb1
Aro80 [3]	Mds3	Stb2
Arr1 [1]	Met18	Stb4
Ash1 [5]	Met28 [3]	Stb5
Ask10	Met31 [3]	Stb6
Azf1	Met32 [3]	Ste12 [5,6]
Bas1 [3]	Met4 [3]	Stp1 [3]
Bye1	Mga1 [1]	Stp2
Cad1 [1,3]	Mig1 [8]	Stp4
Cbf1 [3]	Mig2 [1]	Sum1
Cha4 [3]	Mig3	Sut1
Cin5 [1,2]	Mot3 [1,2,3]	Sut2
Crz1	Msn1	Swi4
Cst6	Msn2 [1,2,4,7,10]	Swi5
Cup9	Msn4 [1,2,4,10]	Swi6
Dal80 [4]	Mss11 [5]	Tbs1
Dal81 [3,4]	Mth1 [8]	Tec1 [5,6]
Dal82 [3,4]	Ndd1	Thi2 [12]
Dat1	Ndt80	Tos8
Dig1 [5,6]	Nnf2	Tye7
Dot6	Nrg1 [1,2]	Uga3 [3,4]
Ecm22	Oaf1	Ume6 [1]
Eds1	Opi1	Upc2
Fap7	Pdc2	Usv1
Fhl1 [1,3,4]	Pdr1 [2]	War1
Fkh1	Pdr3	Wtm1
Fkh2 [1,2]	Phd1 [5]	Wtm2
Fzf1	Pho2 [1,2,3,11]	Xbp1 [2,7]
Gal3	Pho4 [11]	Yap1 [1,2,7]
Gal4 [8,9]	Pip2	Yap3 [1]
Gal80	Ppr1	Yap5 [1]
Gat1 [3,4,7]	Put3 [2,3]	Yap6 [1,2]
Gat3	Rap1 [3]	Yap7 [1,2]
Gcn4 [3,4]	Rco1	YBL054W
Gcr1	Rcs1 [1,2,3]	YBR239C
Gcr2 [3]	Rdr1	YBR267W

Gln3 [3,4]	Rds1 [1]	YDR026C
Gts1	Reb1 [1,2]	YDR049W
Gzf3 [1,4]	Rfx1	YDR266C
Haa1	Rgm1	YDR520C
Hac1	Rgt1 [8]	YER051W
Hal9	Rim101 [1,2]	YER130C
Hap1	Rlm1 [5]	YER184C
Hap2 [4]	Rlr1	YFL044C
Hap3	Rme1	YFL052W
Hap4 [2,3]	Rox1 [1,2]	YGR067C
Hap5 [3]	Rph1 [1,2,3]	Yhp1
Hir1	Rpi1	YJL206C [1,2]
Hir2	Rpn4 [1,2]	YKL222C
Hir3	Rtg1 [3,4]	YKR064W
Hms1	Rtg3 [1,2,3,4]	YLR278C
Hms2	Rts2	YML081W
Hog1	Sfl1	YNR063W
Hsf1 [1,2,7]	Sfp1 [1,2,3]	Yox1
Ifh1	Sig1 [1]	YPR022C
Ime1 [1]	Sip3	YPR196W
Ime4 [1]	Sip4 [3]	Yrr1
Ino2	Skn7 [1,2,7]	Zap1
Ino4	Sko1	Zms1
Ixr1	Smk1	

1 - highly hyperoxic	7 - elevated temperature
2- moderately hyperoxic	8 - galactose medium
3 - amino acid starved	9 - raffinose medium
4 - nutrient deprived	10 - acidic medium
5 - filamentation inducing	11 - phosphate deprived medium
6 - mating inducing	12 - vitamin deprived medium

Supplementary Table S2

Regulator	Discovered Specificity[1]	Known Specificity[1,2]	Programs[3]
Abf1	rTCAyt....Acg	rTCAyT....ACGw	A, C, D, K, M, N
Ace2	tGCTGGT	GCTGGT	K
Adr1		GGrGk	
Aft2	GGGTGy	"ATCTTCAAAAGTGCACCCAT... ...TTGCAGGTGC"	A, C, D, M, N
Arr1		TTACTAA	
Ash1		yTGACT	
Azf1	YwTTkcKkTyyckgykky	AAGAAAAA	N
Bas1		TGACTC	A, K, M, N
Cad1	mTTAsTmAkC	TTACTAA	A, C, D, M, N
Cbf1	tCACGTG	rTCACrTGA	A, C, D, K, M, N
Cin5	TTAygTAA	TTACTAA	A, C, D
Crz1		GwGGCTG	
Dal80	GATAA	GATAAG	
Dal81		AAAAGCCGCGGGCGGGATT	
Dal82	GATAAGa	"GCTGAAAGTTGCGGTGCGATA... ...GAATACCGCGGATTTTGGAA"	K, D
Dig1	TgAAAca		A, C, D, K, M, N
Ecm22		CTCGTATAAGC	
Fhl1	rTGTayGGrtg		A, C, D, K, M, N
Fkh1	tTgTTTac	GGTAAACAA	A, C, D, K, M, N
Fkh2	aaa.GTAAACAa	GGTAAACAA	A, C, D, K, M, N
Gal4	CGG...........cCg	CGG...........CCG	A, K
Gal80		CGG...........CCG	
Gat1	aGATAAG	GATAA	K
Gcn4	TGAsTCa	ArTGACTCw	A, C, D, K, M, N
Gcr1	GGCTTCCwC		
Gln3	GATAAGa.a	GATAAGATAAG	C, D, K
Gzf3	GATAAG	GATAA	
Hac1	kGmCAGCGTGTC		
Hap1	GGmraTA.CGs	CGG...TA.CGG	C, M
Hap2	CCAAT	ayc..ccaat.a.m	
Hap3	CCAAT	ayc..ccaat.a.m	
Hap4	g.CcAAtcA	ayc..ccaat.a.m	A, C, D, M, N
Hap5	CCAAT		
Hsf1	TTCya.....	TTCAGAA..TTCTAGAA	A, C, D, K, M, N
Ime1	AAkGAAA.kwA		A
Ino2	CAcaTGc	GATGTGAAAT	C, D, M, N
Ino4	CATGTGaaaa	CATGTGAAAT	A, C, D, K, M, N
Leu3	cCGgtacCGG	yGCCGGTACCGGyk	A, D, K, M,
Mac1	GAGCAAA		
Mbp1	rACGCGt	ACGCGT	A, C, D, K, M, N

Mcm1	tttCC.rAt..gg	wTTCCyAAw..GGTAA	A, C, D, M, N
Met31		AAACTGTGG	
Met32		AAACTGTGG	
Met4	RMmAwsTGKSgyGsc		C
Mot3		yAGGyA	
Msn2	mAGGGGsgg	mAGGGG	M
Msn4		mAGGGG	
Ndd1	tt.CC.rAw..GG		A, D
Nrg1	GGaCCCT	TCCCTCATTTC	A, C, D, M, N
Opi1	TCGAAyC		
Pdr1	ccGCCgRAwra	CCGCGG	M
Pdr3		TCCGCGGA	
Phd1	sc.GC.gg		A, D, N
Pho2	SGTGCGsygyG		N
Pho4	CACGTGs	cacgtk.g	K, D, N
Put3		CGG..........CCG	
Rap1	tGyayGGrtg	wrmACCCATACAyy	A, C, D, M, N
Rcs1	ggGTGca.t	AmTGCACCCAkTT	C, D, M, N
Rds1	kCGGCCGa		D, N
Reb1	CGGGTAA	TTACCCGG	A, C, D, K, M, N
Rfx1	TTgccATggCAAC		D
Rgt1		CGGA..A	
Rim101		TGCCAAG	
Rlm1		CTAwwwwTAG	
Rlr1	ATTTTCttCwTt		N
Rox1		ysyATTGTT	
Rph1	CCCCTTAAGG	AGGGG	
Rpn4	TTTGCCACCGGTGGCAAA		A, C, D, K, M, N
Rtg1		GGTCAC	
Sfl1		GAAGCTTC	
Sfp1	ayCcrtACay		A, C, D, M, N
Sig1	ArGmAwCrAmAA		M
Sip4	CGG.y.AATGGrr	yCGGAyrrAwGG	D
Skn7	G.C..GsCs	ATTTGGCyGGsCC	A, C, D, M, N
Sko1		ACGTCA	
Smp1		ACTACTAwwwwTAG	
Snt2	yGGCGCTAyca		A, C, D, M, N
Sok2	tGCAg..a		A
Spt2	ymtGTmTytAw		M
Spt23	rAAATsaA		C
Stb1	rracGCsAaa		C, D, K, M, N
Stb4	TCGg..CGA		K
Stb5	CGGwstTAta	CCG	D, N
Ste12	tgAAACa	ATGAAAC	A, C, D, K, M, N
Stp1		rCGGC...rCGGC	
Sum1	gyGwCAswaaw	AGyGwCACAAAAk	A, C, D, M, N
Sut1	gcsGsg..sG		A, D, M
Swi4	raCgCsAAA	C.CGAAA	A, C, D, K, M, N

Swi5		kGCTGr	
Swi6	tttcGCGt	C.CGAAA	A, C, D, M, N
Tec1	rrGAATG	CATTCy	
Thi2	gmAAcy.twAgA		C, D
Tye7	tCACGTGAy	CA..TG	A, C, D, M
Uga3		CCG....CGG	
Ume6	taGCCGCCsa	wGCCGCCGw	A, C, D, K, M, N
Xbp1		CTTCGAG	
Yap1	TTaGTmAGc	TGAsTCAG	A, C, D, M
Yap3		TTACTAA	
Yap5		TTACTAA	
Yap6		TTACTAA	
Yap7	mTkAsTmAk	TTACTAA	A, C, D, M, N
Ydr026C		ttTACCCGGm	C, D, M, N
Yhp1		TAATTG	
Yox1	AsAATA.TGAmr	yAATTA	
Zap1	ACCCTmAAGGTyrT	ACCCTAAAGGT	

1Ambiguity Codes: S = CG, W = AT, R = AG, Y = CT, K = GT, M = AC, "." = ACGT. Letter capitalization reflects the information content at each position of the motif.
2Known specificities are taken from the YPD, SCPD, and TRANSFAC databases.

3Program Codes: A = AlignACE, C = CONVERGE, D = MDscan, K = Kellis et. al, M = MEME, N = MEME (consensus genome)

Supplementary Table OS3

Regulator	Environment	Functional category
Cbf1	SM	amino acid metabolism
Dig1	But	morphogenesis
Fhl1	YPD, SM, Rap,	Protein biosynthesis
Fkh1	YPD	Cell cycle
Gal4	Gal, Raf	carbohydrate metabolism
Gcn4	YPD, SM, Rap	amino acid metabolism
Gln3	YPD	amino acid metabolism
Hap1	YPD	electron transport
Hap2	Rap	cellular respiration
Hap4	YPD	cellular respiration
Hsf1	H2O2Hi, H2O2Lo	response to stress
Ino4	YPD	lipid metabolism
Leu3	YPD	amino acid metabolism
Mbp1	H2O2Hi, H2O2Lo	DNA metabolism
Mcm1	Alpha	Cell cycle
Msn2	H2O2Lo	carbohydrate metabolism
Rcs1	H2O2Hi, H2O2Lo	transport
Rds1	H2O2Hi	carbohydrate metabolism
Reb1	H2O2Lo	vesicle-mediated transport
Rpn4	H2O2Hi, H2O2Lo	protein catabolism
Sok2	But	meiosis
Ste12	YPD, Alpha, But,	conjugation
Sum1	YPD	sporulation
Tec1	But	cell wall organization
Yap6	H2O2Lo	response to stress

Supplemental Table S4

Factor	Pvalue	Single	Repetitive	Likely Category	Motif	Bits
SKN7	9.85E-01	O: 41 E: 40	O: 16 E: 16	-	G.C..GsCs (1)	6.0339
MOT3	8.42E-01	O: 6 E: 5	O: 2 E: 2	-	yAGGyA (1)	8.0627
HAC1	5.31E-01	O: 1 E: 0	O: 0 E: 0	-	kGmCAGCGTGTC (0)	18.8864
ROX1	1.59E-01	O: 10 E: 7	O: 1 E: 3	-	ysyATTGTT (0)	12.8397
YAP1	1.26E-01	O: 11 E: 8	O: 1 E: 3	-	TTaGTmAGc (1)	11.5264
RAP1	1.48E-01	O: 34 E: 38	O: 20 E: 15	-	tGyayGGrtg (1)	9.3829
GAL80	1.10E-01	O: 0 E: 0	O: 1 E: 0	-	CGG..........CCG (1)	10.3267
HAP2	5.74E-01	O: 14 E: 12	O: 4 E: 5	-	CCAAT (0)	8.6055
INO2	4.77E-02	O: 11 E: 15	O: 10 E: 5	Repetitive	CAcaTGc (1)	8.8175
HAP1	4.74E-03	O: 31 E: 23	O: 2 E: 9	Single	GGmraTA.CGs (1)	11.2364
HAP4	2.10E-01	O: 20 E: 17	O: 4 E: 6	-	g.CcAAtcA (1)	7.2832
INO4	4.90E-01	O: 15 E: 13	O: 4 E: 5	-	CATGTGaaaa (1)	10.8029
RLM1	6.02E-02	O: 9 E: 6	O: 0 E: 2	-	CTAwwwwTAG (0)	13.6774
MET31	8.88E-01	O: 3 E: 2	O: 1 E: 1	-	AAACTGTGG (0)	15.49
MET32	5.65E-01	O: 8 E: 7	O: 2 E: 2	-	AAACTGTGG (0)	15.49
MCM1	4.58E-01	O: 28 E: 30	O: 14 E: 11	-	tttCC.rAt..gg (1)	7.8147
LEU3	6.91E-01	O: 7 E: 6	O: 2 E: 2	-	cCGgtacCGG (1)	10.3977
SWI6	7.36E-06	O: 34 E: 50	O: 37 E: 20	Repetitive	tttcGCGt (1)	7.33
AZF1	5.31E-01	O: 1 E: 0	O: 0 E: 0	-	YwTTkcKkTyyckgykky (1)	15.7085
RIM101	5.31E-01	O: 1 E: 0	O: 0 E: 0	-	TGCCAAG (0)	12.0478
GAT1	8.85E-02	O: 3 E: 5	O: 4 E: 1	-	aGATAAG (1)	10.571
IME1	2.78E-01	O: 3 E: 2	O: 0 E: 0	-	AAkGAAA.kwA (1)	13.1852
UME6	6.31E-03	O: 55 E: 45	O: 8 E: 17	Single	taGCCGCCsa (1)	10.4956
MBP1	2.99E-08	O: 34 E: 56	O: 44 E: 21	Repetitive	rACGCGt (1)	8.8321
PHO4	8.96E-01	O: 11 E: 10	O: 4 E: 4	-	CACGTGs (1)	8.9549
PHO2	2.78E-01	O: 3 E: 2	O: 0 E: 0	-	SGTGCGsygyG (1)	13.7626

30

STB4	6.84E-01 O: 4 E: 3	O: 1 E: 1	-	TCGg..CGA (1)	10.3574
STB5	3.75E-01 O: 10 E: 8	O: 2 E: 3	-	CGGwstTAta (1)	10.3136
STB1	9.29E-02 O: 6 E: 8	O: 6 E: 3	-	rracGCsAaa (1)	7.7461
NRG1	1.63E-01 O: 35 E: 30	O: 8 E: 12	-	GGaCCCT (1)	10.1382
GZF3	2.78E-01 O: 3 E: 2	O: 0 E: 0	-	GATAAG (0)	10.3267
SFP1	3.03E-02 O: 7 E: 10	O: 8 E: 4	Repetitive	ayCcrtACay (1)	7.8755
YHP1	3.76E-01 O: 2 E: 1	O: 0 E: 0	-	TAATTG (0)	10.3267
DIG1	1.43E-08 O: 25 E: 45	O: 38 E: 17	Repetitive	TgAAAca (1)	7.5967
MET4	1.61E-01 O: 5 E: 3	O: 0 E: 1	-	RMmAwsTGKSgyGsc (1)	16.622
NDD1	8.78E-02 O: 21 E: 17	O: 3 E: 6	-	tt.CC.rAw..GG (1)	10.2085
MAC1	8.88E-01 O: 3 E: 2	O: 1 E: 1	-	GAGCAAA (0)	12.0478
SNT2	8.02E-02 O: 13 E: 10	O: 1 E: 3	-	yGGCGCTAyca (1)	10.9134
ADR1	8.98E-01 O: 7 E: 7	O: 3 E: 2	-	GGrGk (0)	6.8387
XBP1	3.76E-01 O: 2 E: 1	O: 0 E: 0	-	CTTCGAG (0)	12.0478
ASH1	5.31E-01 O: 1 E: 0	O: 0 E: 0	-	yTGACT (0)	9.4432
RPH1	5.31E-01 O: 1 E: 0	O: 0 E: 0	-	CCCCTTAAGG (0)	17.2111
PDR3	1.10E-01 O: 0 E: 0	O: 1 E: 0	-	TCCGCGGA (0)	13.7689
STE12	5.57E-04 O: 48 E: 62	O: 39 E: 24	Repetitive	tgAAACa (1)	7.7973
RCS1	4.01E-01 O: 12 E: 13	O: 7 E: 5	-	ggGTGca.t (1)	7.287
TYE7	4.24E-01 O: 19 E: 17	O: 5 E: 6	-	tCACGTGAy (1)	10.327
ZAP1	3.76E-01 O: 2 E: 1	O: 0 E: 0	-	ACCCTmAAGGTyrT (1)	21.8082
HSF1	1.76E-02 O: 49 E: 40	O: 8 E: 16	Single	TTCya.....TTC (1)	8.4321
YDR026c	4.76E-02 O: 10 E: 7	O: 0 E: 2	Single	ttTACCCGGm (1)	15.1416
SPT2	2.01E-01 O: 9 E: 7	O: 1 E: 2	-	ymtGTmTytAw (1)	11.107
SIG1	3.76E-01 O: 2 E: 1	O: 0 E: 0	-	ArGmAwCrAmAA (1)	14.1436
SWI4	7.29E-04 O: 27 E: 38	O: 26 E: 14	Repetitive	raCgCsAAA (1)	9.5841
GLN3	3.01E-01 O: 19 E: 21	O: 11 E: 8	-	GATAAGa.a (1)	9.7858

31

ABF1	1.68E-04 O: 119 E: 99	O: 19 E: 38	Single	rTCAyt....Acg (1)	8.7986
PHD1	7.89E-03 O: 15 E: 21	O: 15 E: 8	Repetitive	sc.GC.gg (1)	6.6163
STP1	5.31E-01 O: 1 E: 0	O: 0 E: 0	-	rCGGC...rCGGC (1)	15.4442
SUM1	3.56E-01 O: 22 E: 24	O: 12 E: 9	-	gyGwCAswaaw (1)	7.5512
RFX1	4.14E-01 O: 6 E: 5	O: 1 E: 1	-	TTgccATggCAAC (1)	14.949
AFT2	9.73E-03 O: 22 E: 29	O: 19 E: 11	Repetitive	GGGTGy (1)	6.2628
RGT1	5.69E-03 O: 0 E: 2	O: 3 E: 0	-	CGGA..A (1)	8.6055
PDR1	3.76E-01 O: 2 E: 1	O: 0 E: 0	-	ccGCCgRAwra (1)	11.0187
YAP3	5.31E-01 O: 1 E: 0	O: 0 E: 0	-	TTACTAA (0)	12.0478
YAP5	5.31E-01 O: 1 E: 0	O: 0 E: 0	-	TTACTAA (0)	12.0478
YAP7	4.74E-02 O: 43 E: 36	O: 8 E: 14	Single	mTkAsTmAk (1)	8.554
YAP6	2.78E-01 O: 3 E: 2	O: 0 E: 0	-	TTACTAA (0)	12.0478
GCN4	9.62E-01 O: 67 E: 66	O: 26 E: 26	-	TGAsTCa (1)	7.7817
OPI1	3.76E-01 O: 2 E: 1	O: 0 E: 0	-	TCGAAyC (0)	11.1643
TEC1	5.02E-02 O: 18 E: 22	O: 14 E: 9	-	rrGAATG (1)	8.4134
YOX1	2.10E-01 O: 4 E: 2	O: 0 E: 1	-	AsAATA.TGAmr (1)	16.2819
GCR1	3.32E-01 O: 2 E: 2	O: 2 E: 1	-	GGCTTCCwC (0)	14.6065
CAD1	6.02E-02 O: 9 E: 6	O: 0 E: 2	-	mTTAsTmAkC (1)	12.0785
RLR1	2.10E-01 O: 4 E: 2	O: 0 E: 1	-	ATTTTCttCwTt (1)	13.9257
CBF1	2.49E-03 O: 87 E: 73	O: 15 E: 28	Single	tCACGTG (1)	10.7486
FKH2	4.13E-01 O: 40 E: 37	O: 12 E: 14	-	aaa.GTAAACAa (1)	10.6377
SKO1	4.14E-01 O: 6 E: 5	O: 1 E: 1	-	ACGTCA (0)	10.3267
HAP3	3.75E-01 O: 10 E: 8	O: 2 E: 3	-	CCAAT (0)	8.6055
FKH1	6.66E-01 O: 43 E: 44	O: 19 E: 17	-	tTgTTTac (1)	8.864
HAP5	6.70E-01 O: 13 E: 12	O: 4 E: 4	-	CCAAT (0)	8.6055
RDS1	4.14E-01 O: 6 E: 5	O: 1 E: 1	-	kCGGCCGa (1)	10.3555
MSN4	5.74E-01 O: 14 E: 12	O: 4 E: 5	-	mAGGGG (0)	9.4432

32

FHL1	5.26E-02 O: 37 E: 43	O: 24 E: 17	-	rTGTayGGrtg (1)	10.6834
SIP4	1.38E-01 O: 1 E: 2	O: 2 E: 0	-	CGG.y.AATGGrr (1)	13.221
ARR1	5.31E-01 O: 1 E: 0	O: 0 E: 0	-	TTACTAA (0)	12.0478
BAS1	2.84E-04 O: 6 E: 12	O: 12 E: 5	Repetitive	TGACTC (1)	10.093
REB1	1.24E-05 O: 135 E: 110	O: 19 E: 43	Single	CGGGTAA (1)	9.5418
SWI5	1.05E-02 O: 11 E: 16	O: 12 E: 6	Repetitive	kGCTGr (0)	8.5598
GAL4	6.91E-01 O: 7 E: 6	O: 2 E: 2	-	CGG...........cCg (1)	6.7079
DAL81	5.31E-01 O: 1 E: 0	O: 0 E: 0	-	AAAAGCCGCGGGCGGGATT (0)	32.7011
DAL80	3.24E-01 O: 7 E: 5	O: 1 E: 2	-	GATAA (0)	8.6055
DAL82	9.45E-02 O: 11 E: 14	O: 9 E: 5	-	GATAAGa (1)	9.8108
CIN5	3.31E-01 O: 42 E: 38	O: 12 E: 15	-	TTAygTAA (1)	8.7034
RTG3	1.61E-01 O: 5 E: 3	O: 0 E: 1	-	GGTCAC (0)	10.3267
SMP1	3.76E-01 O: 2 E: 1	O: 0 E: 0	-	ACTACTAwwwwTAG (0)	20.5618
SOK2	1.34E-05 O: 13 E: 24	O: 21 E: 9	Repetitive	tGCAg..a (1)	6.2038
UGA3	5.31E-01 O: 1 E: 0	O: 0 E: 0	-	CCG....CGG (1)	10.3267
SPT23	2.30E-01 O: 16 E: 13	O: 3 E: 5	-	rAAATsaA (1)	9.171
THI2	1.25E-01 O: 6 E: 4	O: 0 E: 1	-	gmAAcy.twAgA (1)	9.9758
PUT3	8.43E-01 O: 2 E: 2	O: 1 E: 0	-	CGG..........CCG (1)	10.3267
SUT1	7.85E-01 O: 11 E: 11	O: 5 E: 4	-	gcsGsg..sG (1)	6.5317
RPN4	4.01E-04 O: 43 E: 32	O: 2 E: 12	Single	TTTGCCACC (1)	13.3079
ACE2	8.85E-02 O: 3 E: 5	O: 4 E: 1	-	tGCTGGT (1)	10.4997
MSN2	1.53E-02 O: 15 E: 10	O: 0 E: 4	Single	mAGGGGsgg (1)	11.0666

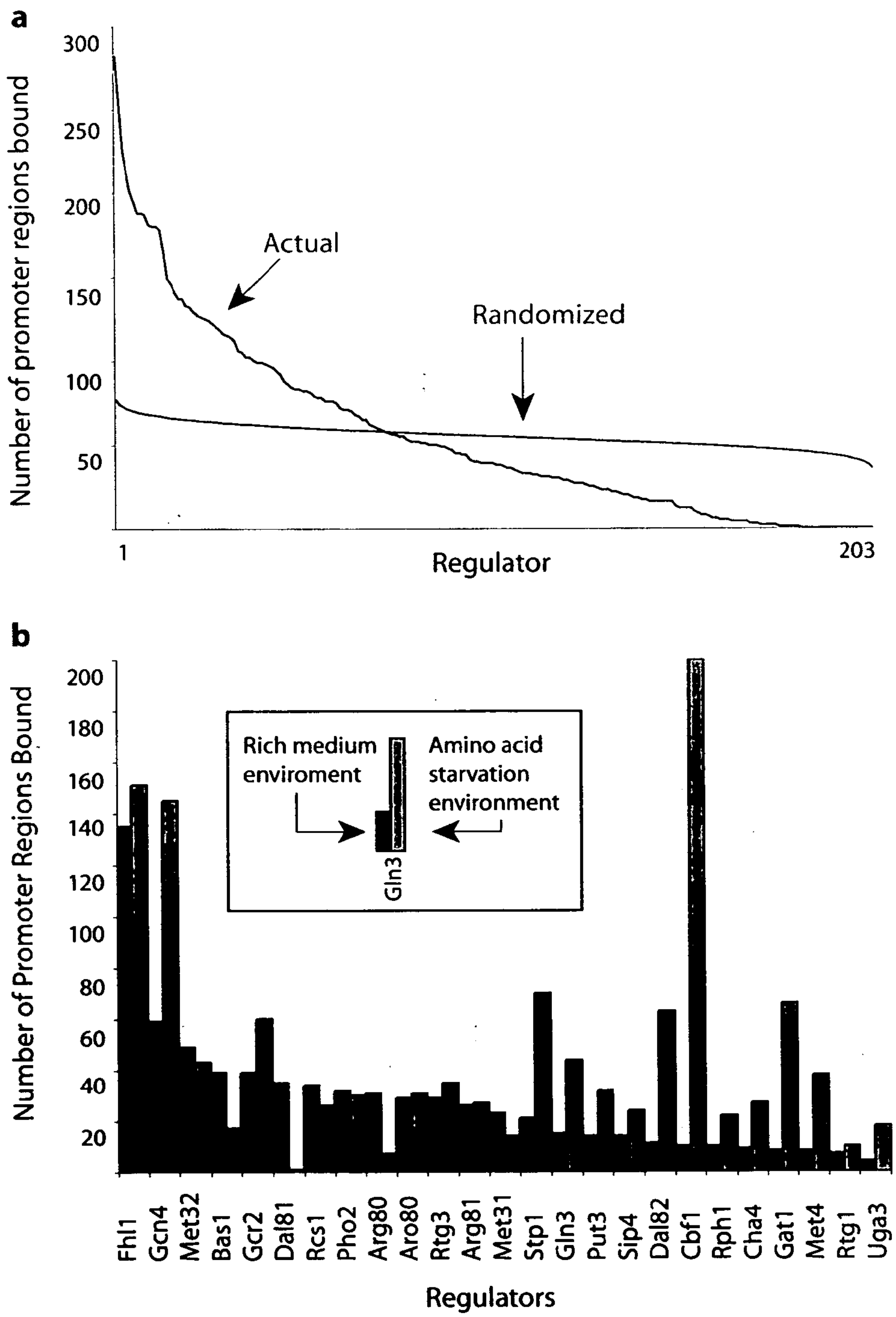
Supplementary Table S5

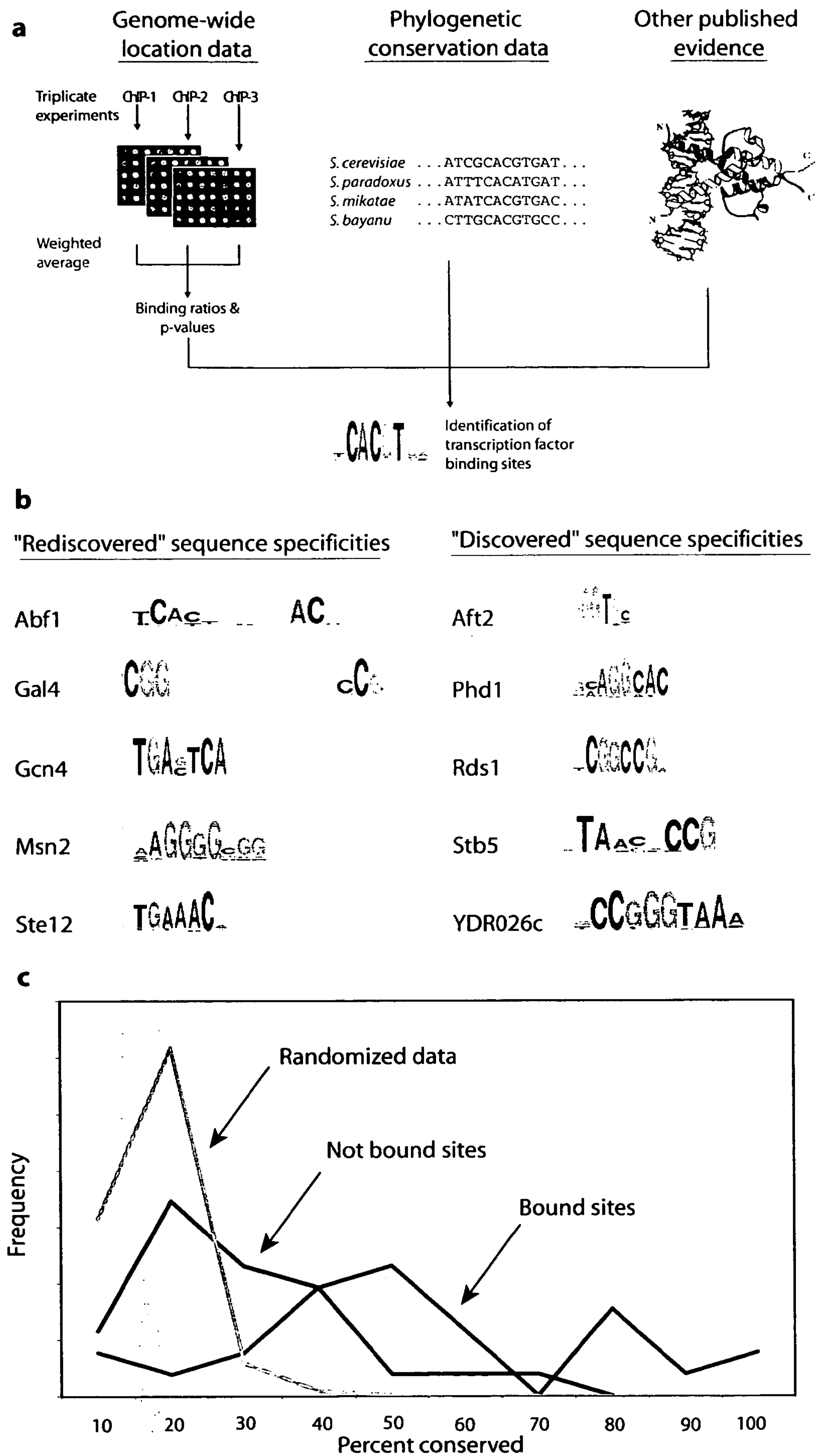
Regulator 1	Regulator 2	P value	Bound by Regulator 1	Bound by Regulator 2	Bound by both
ACE2	SWI5	0	11	23	5
AFT2	RCS1	0	41	19	11
ARR1	YAP3	0.001	1	1	1
AZF1	GZF3	0.002	1	3	1
BAS1	MET4	0.002	18	5	2
CAD1	YAP1	0	9	12	4
CAD1	YAP7	0	9	51	9
CBF1	MET31	0.002	102	4	3
CBF1	MET32	0	102	10	6
CBF1	MET4	0.004	102	5	3
CBF1	PHO4	0.001	102	15	6
CBF1	TYE7	0	102	24	17
CIN5	PHD1	0	54	30	8
CIN5	SKN7	0	54	57	9
CIN5	SOK2	0	54	34	9
CIN5	SUT1	0	54	26	6
CIN5	XBP1	0.002	54	2	2
CIN5	YAP6	0.005	54	3	2
DAL82	GAT1	0	20	7	3
DAL82	GLN3	0	20	30	7
DAL82	HAP2	0.002	20	18	3
DIG1	MCM1	0	63	42	9
DIG1	STE12	0	63	87	49
DIG1	SWI4	0	63	53	14
DIG1	SWI6	0	63	71	12
DIG1	TEC1	0	63	32	16

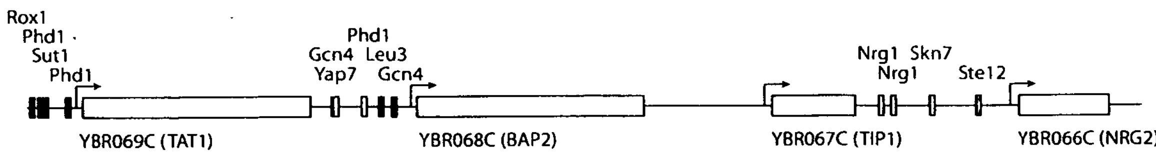
FHL1	RAP1	0	61	54	30
FHL1	SFP1	0	61	15	13
FKH1	FKH2	0	62	52	20
FKH2	MCM1	0	52	42	12
FKH2	NDD1	0	52	24	14
FKH2	SWI6	0	52	71	12
GCN4	GLN3	0.004	93	30	7
GCN4	LEU3	0.002	93	9	4
GCR1	TYE7	0.002	4	24	2
GLN3	HAP2	0	30	18	5
GZF3	PDR1	0.005	3	2	1
HAP2	HAP3	0	18	12	7
HAP2	HAP4	0	18	24	5
HAP2	HAP5	0	18	17	10
HAP3	HAP5	0	12	17	6
HAP4	HAP5	0	24	17	4
HSF1	MSN4	0.001	57	18	5
INO2	INO4	0	21	19	9
INO4	SKO1	0.004	19	7	2
MAC1	RCS1	0.001	4	19	2
MAC1	SUT1	0.002	4	26	2
MBP1	STB1	0	78	12	8
MBP1	SWI4	0	78	53	28
MBP1	SWI6	0	78	71	36
MCM1	NDD1	0	42	24	15
MCM1	STE12	0.001	42	87	9
MCM1	SWI4	0.001	42	53	7
MCM1	SWI6	0	42	71	9
MCM1	TEC1	0.003	42	32	5

MET31	MET32	0	4	10	4
MET31	MET4	0	4	5	4
MET32	MET4	0	10	5	4
MOT3	ROX1	0.002	8	11	2
MOT3	SKN7	0.004	8	57	3
MSN2	MSN4	0	15	18	4
MSN4	NRG1	0.002	18	43	4
NRG1	RLM1	0.002	43	9	3
NRG1	SKN7	0	43	57	9
NRG1	SUT1	0.001	43	26	5
OPI1	PDC2	0	2	2	2
OPI1	YML081W	0	2	2	2
PDC2	YML081W	0	2	2	2
PHD1	ROX1	0	30	11	4
PHD1	SKN7	0	30	57	12
PHD1	SOK2	0	30	34	15
PHD1	SUT1	0	30	26	11
PHD1	SWI6	0	30	71	8
RAP1	SFP1	0	54	15	9
RIM101	YOX1	0.003	1	4	1
RLM1	SKO1	0.001	9	7	2
RLR1	SPT2	0	12	10	4
ROX1	SUT1	0	11	26	4
SIP4	STP1	0.002	3	1	1
SKN7	SOK2	0	57	34	12
SKN7	SUT1	0	57	26	10
SKN7	SWI6	0	57	71	11
SKN7	XBP1	0.002	57	2	2
SKO1	SOK2	0.001	7	34	3

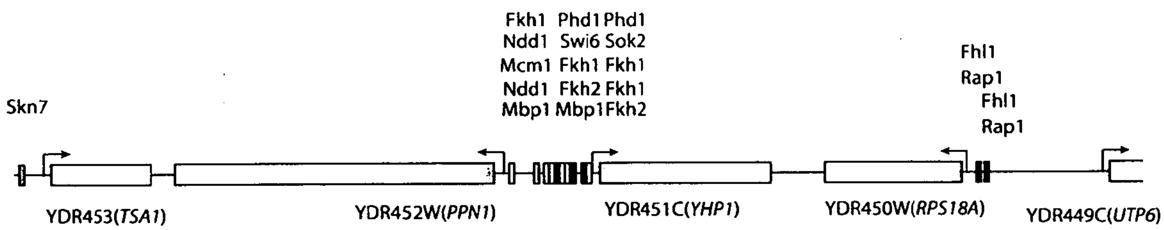
SOK2	SUT1	0	34	26	10
SOK2	SWI6	0.002	34	71	7
STB1	SWI4	0	12	53	10
STB1	SWI6	0	12	71	10
STB1	TEC1	0.003	12	32	3
STE12	SWI4	0	87	53	15
STE12	SWI6	0	87	71	16
STE12	TEC1	0	87	32	18
SWI4	SWI6	0	53	71	43
SWI4	TEC1	0	53	32	11
SWI6	TEC1	0	71	32	11
YAP1	YAP7	0	12	51	8
YAP6	YAP7	0.004	3	51	2


a
300
250
200
150
100
50
Number of promoter regions bound
Actual
Randomized
1
203
Regulator
b
200
180
160
140
120
100
80
60
40
20
Number of Promoter Regions Bound
Rich medium enviroment
Amino acid starvation environment
Gln3
Fhl1
Gcn4
Met32
Bas1
Gcr2
Dal81
Rcs1
Pho2
Arg80
Aro80
Rtg3
Arg81
Met31
Stp1
Gln3
Put3
Sip4
Dal82
Cbf1
Rph1
Cha4
Gat1
Met4
Rtg1
Uga3
Regulators

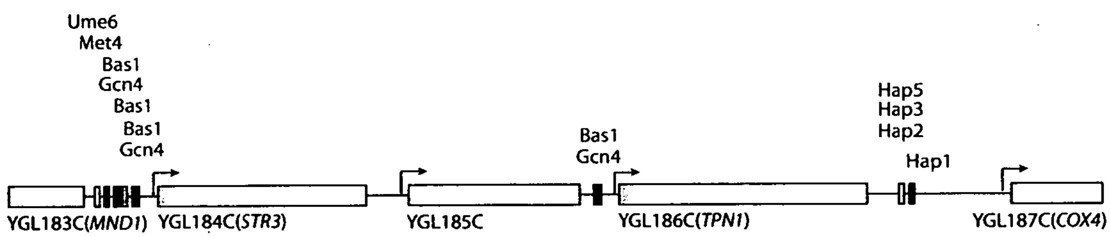

a
Genome-wide location data
Phylogenetic conservation data
Other published evidence
Triplicate experiments
CHIP-1 CHIP-2 CHIP-3
Weighted average
Binding ratios & p-values
S. cerevisiae . . . ATCGCACGTGAT . . .
S. paradoxus . . . ATTTCACATGAT . . .
S. mikatae . . . ATATCACGTGAC . . .
S. bayanu . . . CTTGCACGTGCC . . .
Identification of transcription factor binding sites
b
"Rediscovered" sequence specificities
"Discovered" sequence specificities
Abf1
Gal4
Gcn4
Msn2
Ste12
Aft2
Phd1
Rds1
Stb5
YDR026c
c
Randomized data
Not bound sites
Bound sites
Frequency
10 20 30 40 50 60 70 80 90 100
Percent conserved



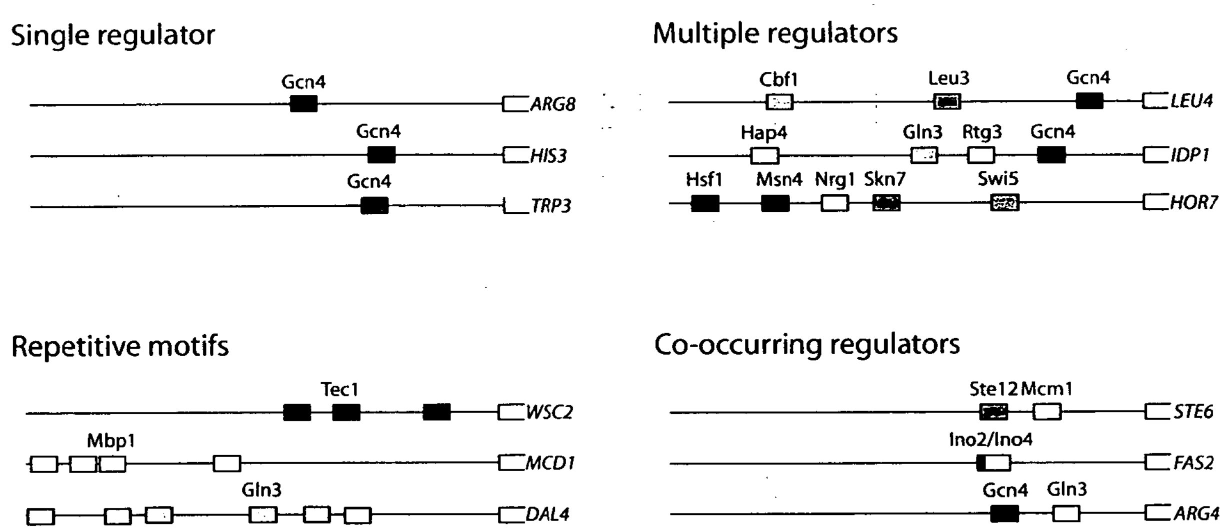
Chromosome II
(370000:379000)

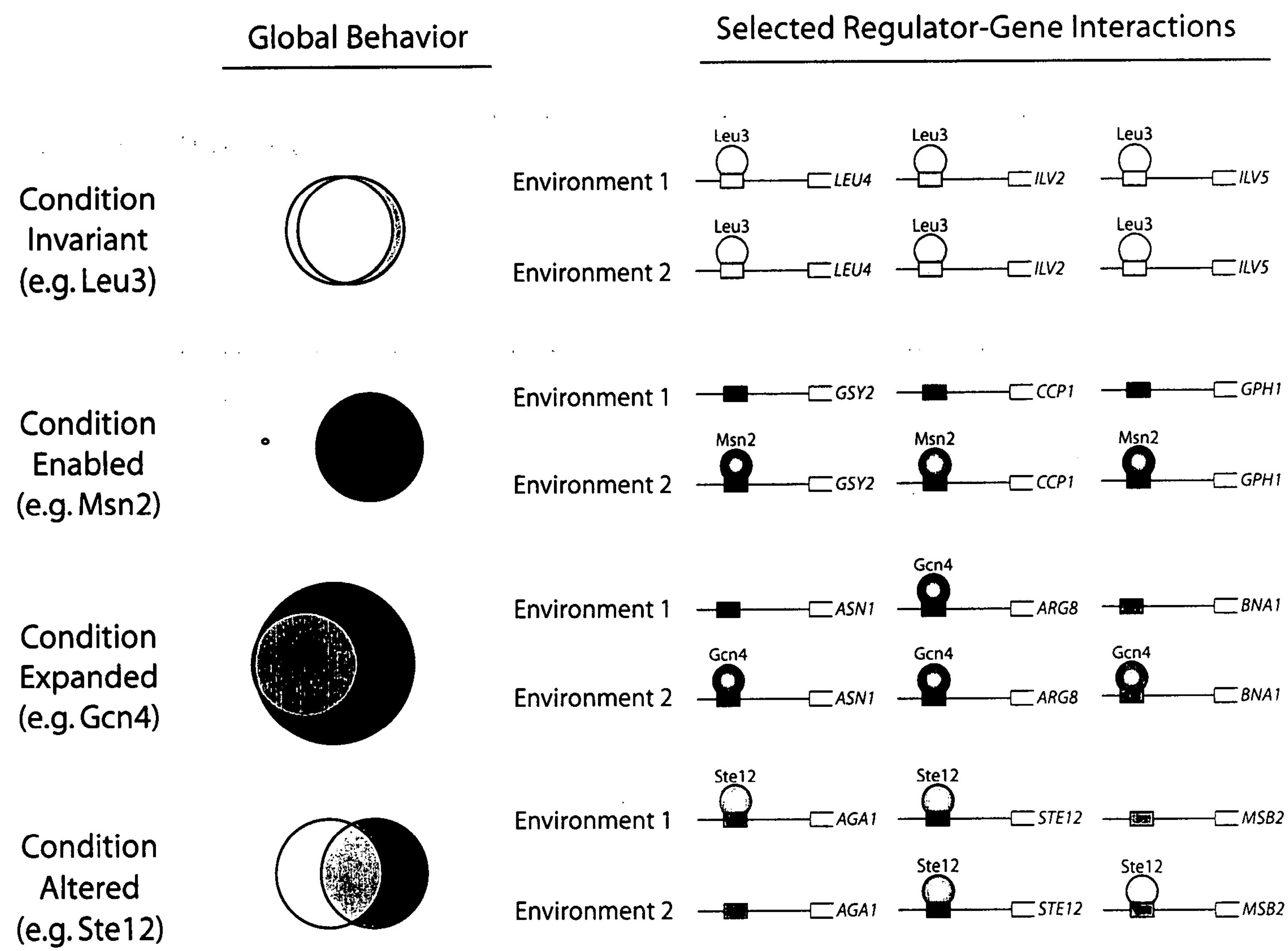


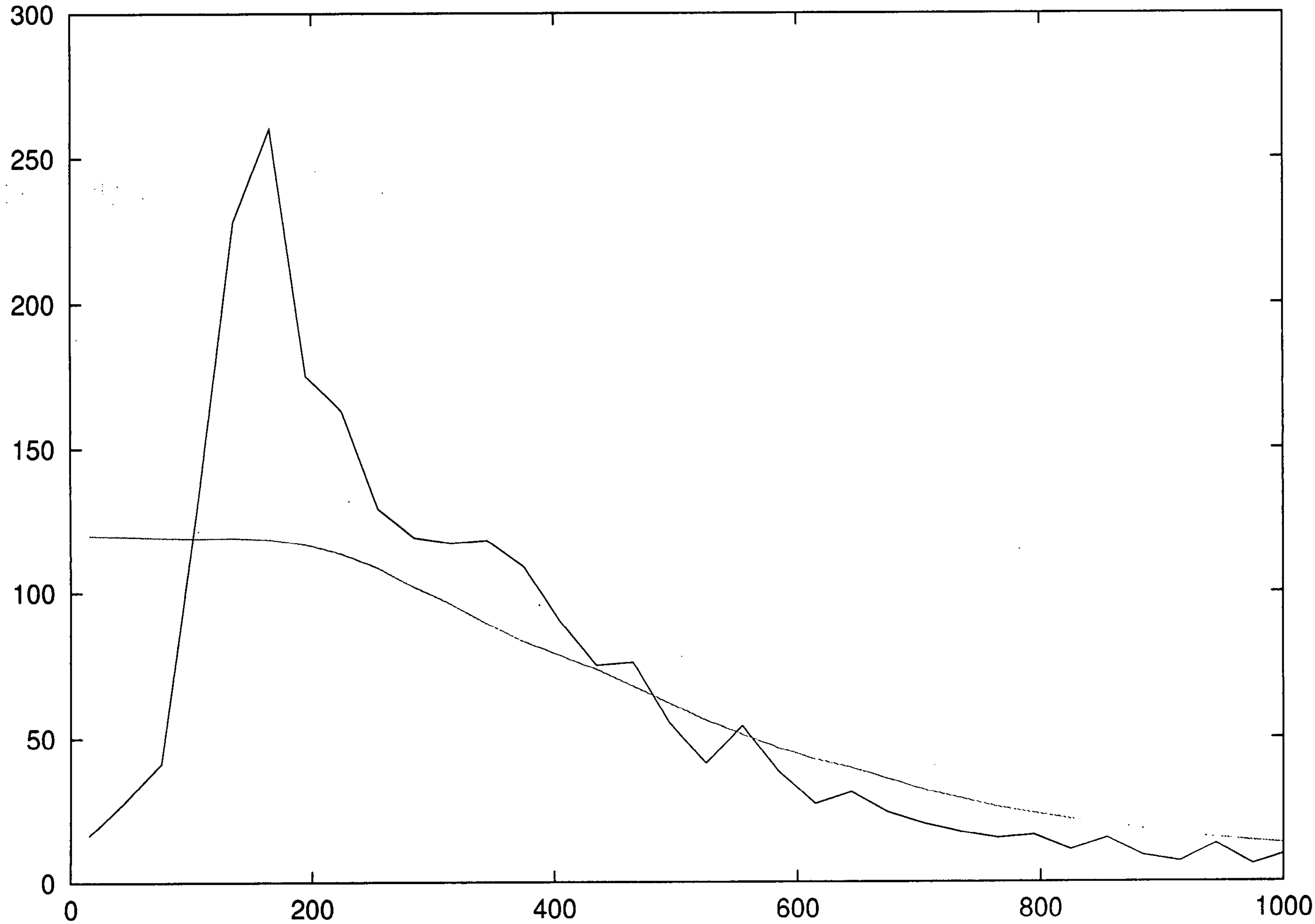
Chromosome IV
(1358800:1366600)



Chromosome V
(150000:157000)

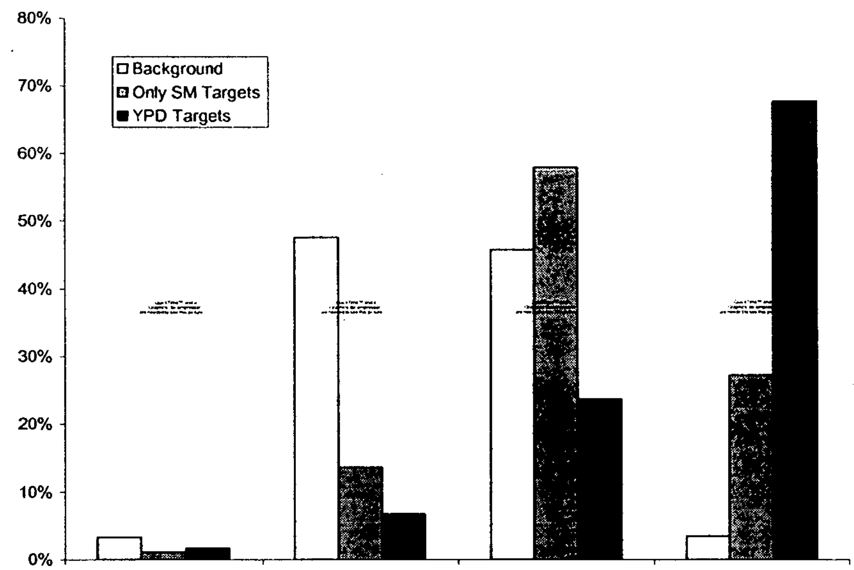

Single regulator
Gcn4
ARG8
Gcn4
HIS3
Gcn4
TRP3
Multiple regulators
Cbf1
Leu3
Gcn4
LEU4
Hap4
Gln3
Rtg3
Gcn4
IDP1
Hsf1
Msn4
Nrg1
Skn7
Swi5
HOR7
Repetitive motifs
Tec1
WSC2
Mbp1
MCD1
Gln3
DAL4
Co-occurring regulators
Ste12
Mcm1
STE6
Ino2/Ino4
FAS2
Gcn4
Gln3
ARG4


Global Behavior
Selected Regulator-Gene Interactions
Condition Invariant (e.g. Leu3)
Environment 1
Leu3
LEU4
Leu3
ILV2
Leu3
ILV5
Environment 2
Leu3
LEU4
Leu3
ILV2
Leu3
ILV5
Condition Enabled (e.g. Msn2)
Environment 1
GSY2
CCP1
GPH1
Environment 2
Msn2
GSY2
Msn2
CCP1
Msn2
GPH1
Condition Expanded (e.g. Gcn4)
Environment 1
ASN1
Gcn4
ARG8
BNA1
Environment 2
Gcn4
ASN1
Gcn4
ARG8
Gcn4
BNA1
Condition Altered (e.g. Ste12)
Environment 1
Ste12
AGA1
Ste12
STE12
MSB2
Environment 2
AGA1
Ste12
STE12
Ste12
MSB2


300
250
200
150
100
50
0
0
200
400
600
800
1000

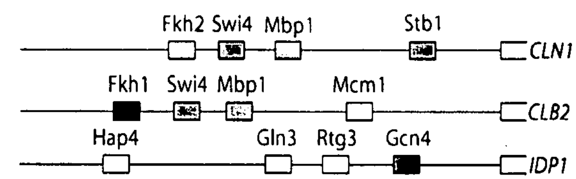
Supplemenary Figure S2

Binding of GCN4 to imperfect sites under stress



Score [% maximum] of matching sites	0%	0%-33%	33%-66%	66%-100%
Consensus of matching sites for each score				

A Draft Transcriptional Regulatory Code for Yeast



Introduction
- Concept
- History

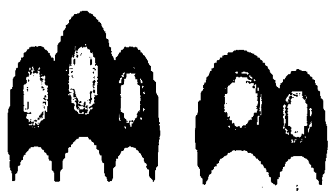
Data and Regulatory Code Assembly
- Genome-wide binding data
- DNA sequence specificities

Promoter Architectures

Environment-dependent Use of Code

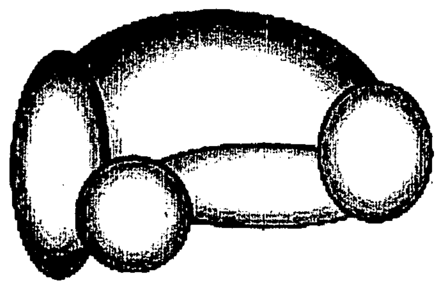
Regulation of Gene Expression

DNA-binding Regulators



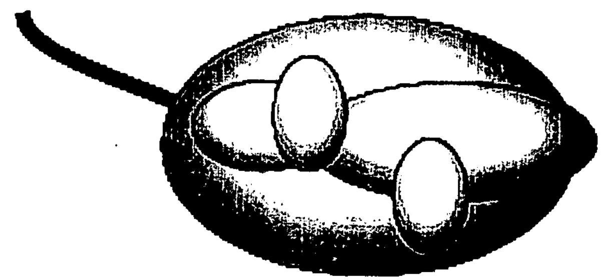
Recognize specific sequences

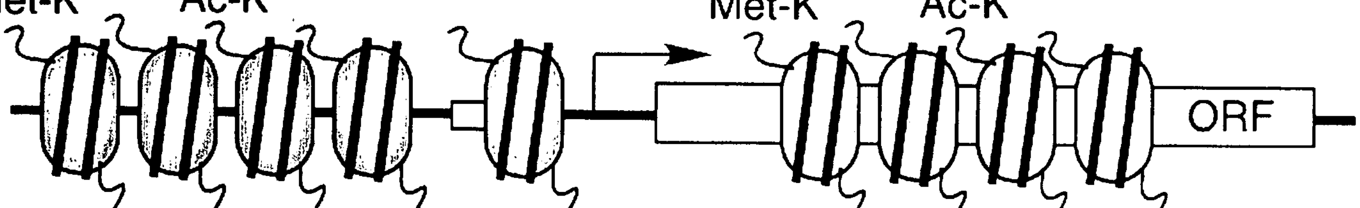
Chromatin Regulators



Control access to sequence

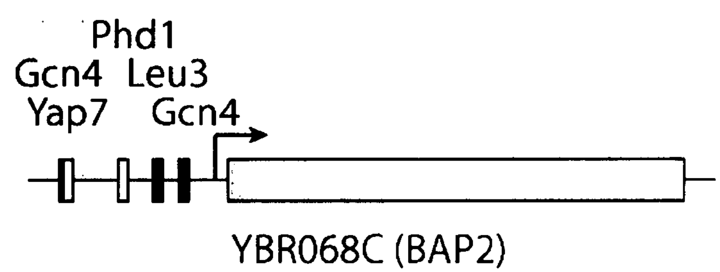
Transcription
Initiation Apparatus





We have considerable knowledge of molecular mechanisms that operate at individual genes, but limited knowledge of global regulatory mechanisms.

Transcriptional Regulatory Code Information



What regulators contribute to control of each gene?

What sequences do they bind?

When do the regulators bind these sequences?

Previous Methods for Identifying Regulatory Regions

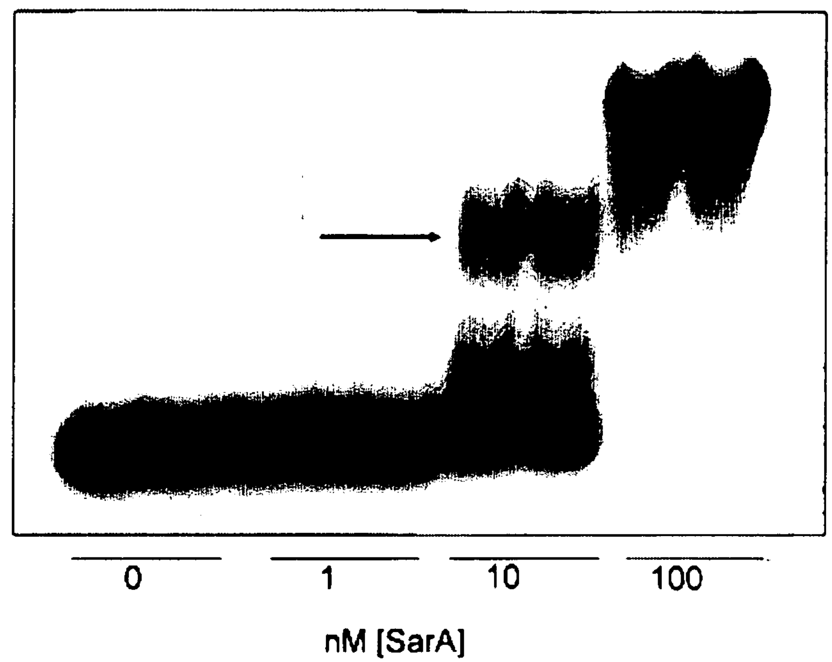
Genetic

Genotype	Beta-Galactosidase	
	Noninduced	Induced
$O^{+} Z^{+} Y^{+}$	-	+
$O^{+} Z^{+} Y^{+}/FO^{+} Z^{-} Y^{+}$	-	+
$O^{c} Z^{+} Y^{+}$	+	+
$O^{+} Z^{+} Y^{-}/FO^{c} Z^{+} Y^{+}$	+	+
$O^{+} Z^{+} Y^{+}/FO^{c} Z^{-} Y^{+}$	-	+
$O^{+} Z^{-} Y^{+}/FO^{c} Z^{+} Y^{-}$	+	+

Previous Methods for Identifying Regulatory Regions

Genetic

Biochemical

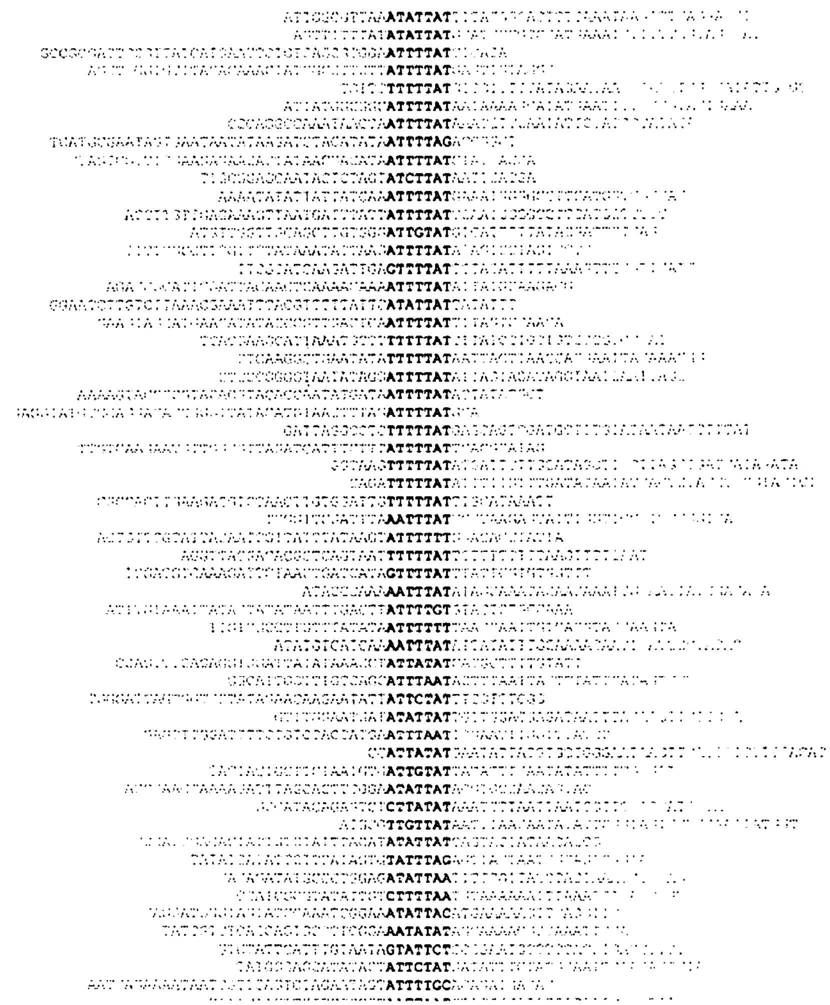


Previous Methods for Identifying Regulatory Regions

Genetic

Biochemical

Alignment



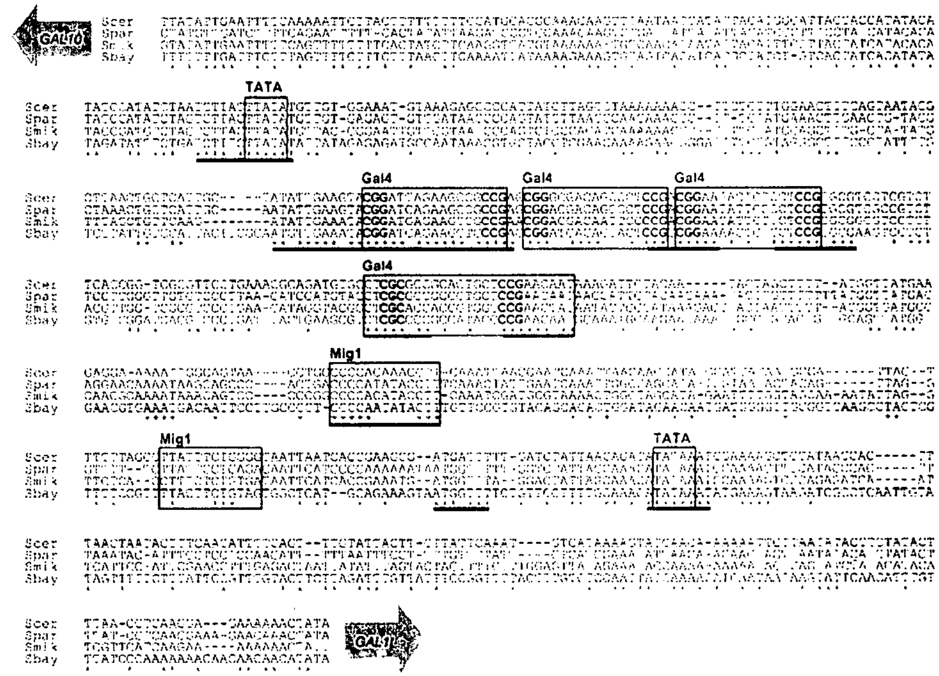
Previous Methods for Identifying Regulatory Regions

Genetic

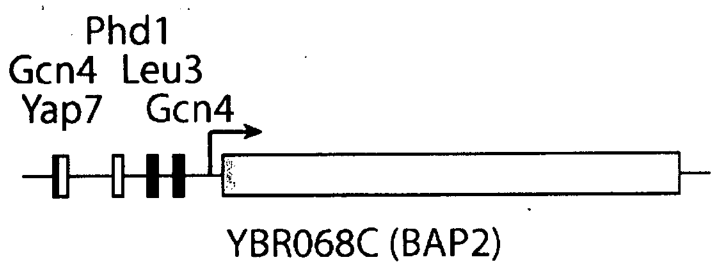
Biochemical

Alignment

Comparative Genomics



Transcriptional Regulatory Code Information

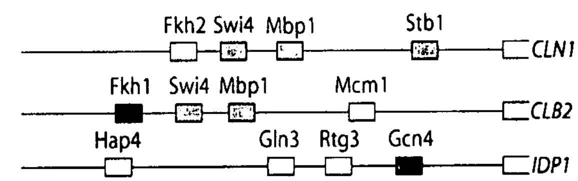


What regulators contribute to control of each gene?

What sequences do they bind?

When do the regulators bind these sequences?

A Draft Transcriptional Regulatory Code for Yeast



Introduction
- Concept
- History

Data and Regulatory Code Assembly
- Genome-wide binding data
- DNA sequence specificities

Promoter Architectures

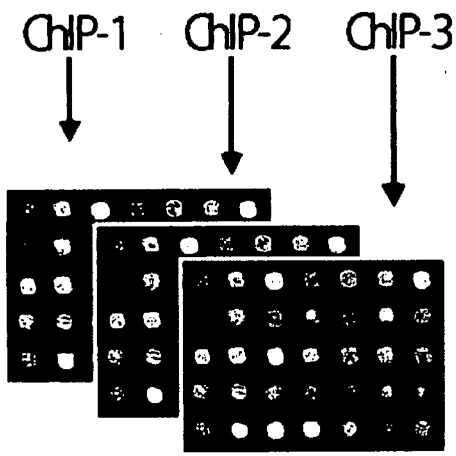
Environment-dependent Use of Code

Combining Information to Discover Regulatory Code

Phylogenetic conservation data

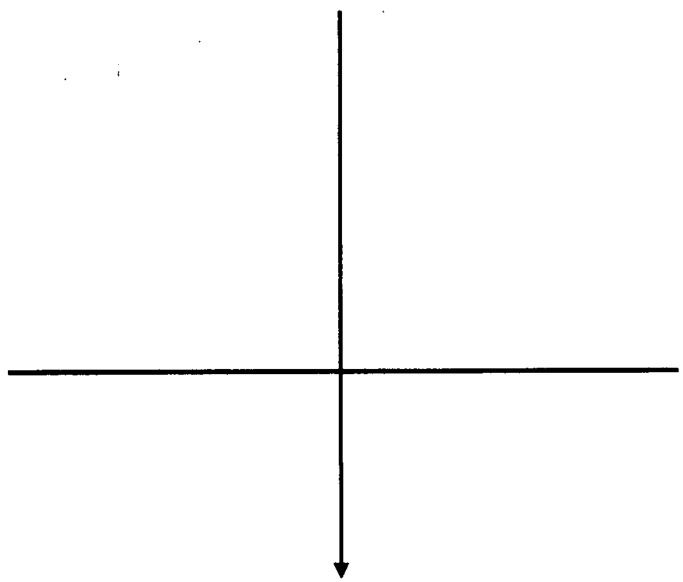
S. cerevisiae ...ATCGCACGTGAT...
S. paradoxus ...ATTTCACATGAT...
S. mikatae ...ATATCACGTGAC...
S. bayanu ...CTTGCACGTGCC...

Genome wide binding data



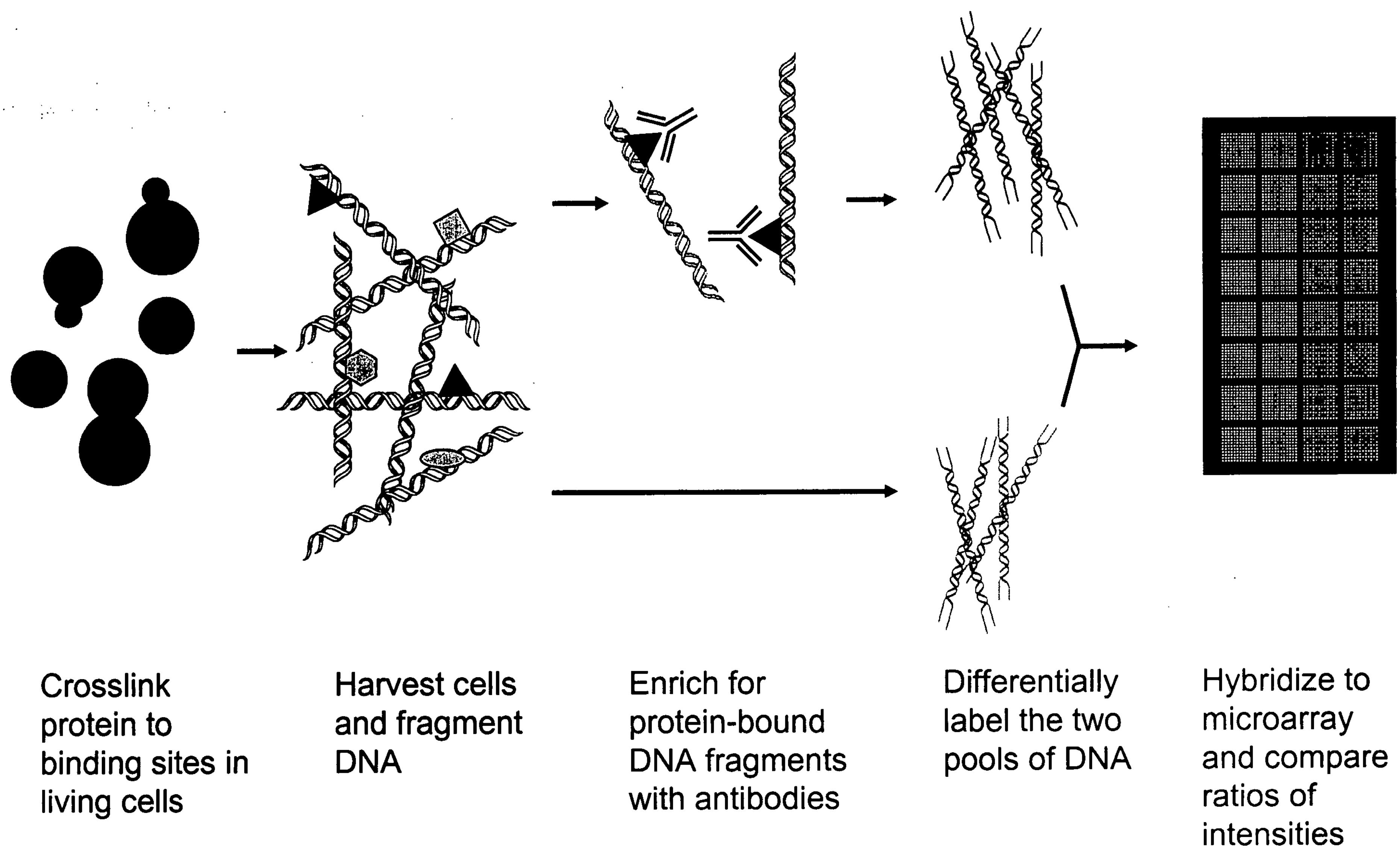
Weighted average:
Binding ratios and
P values

Other published evidence

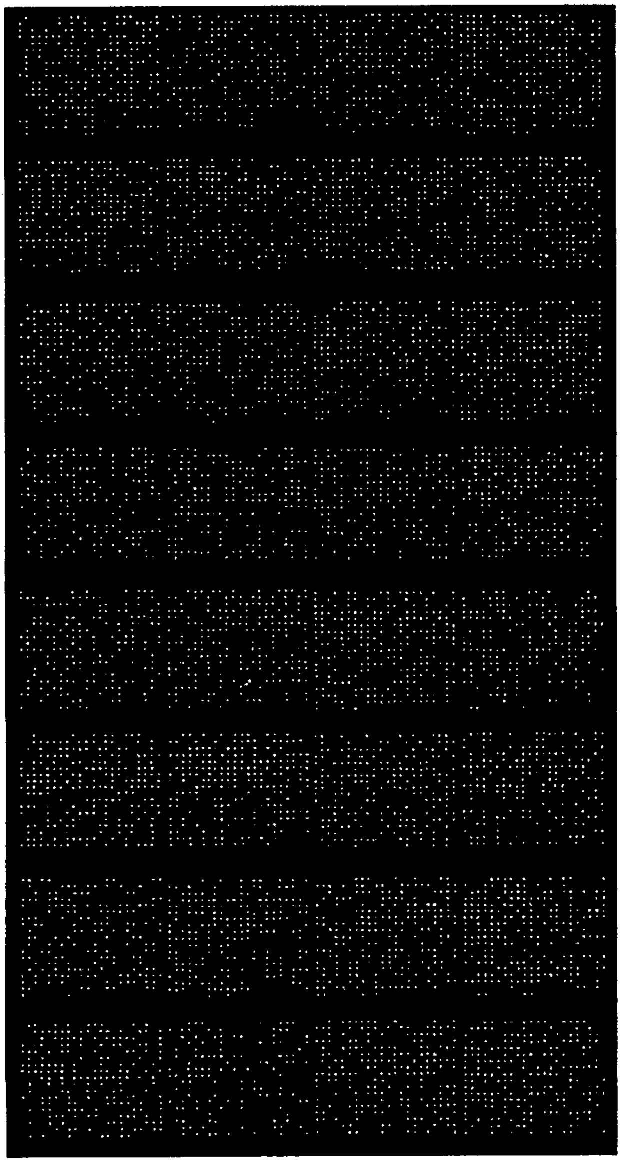


Identification of
transcription factor
binding sites

Genome-wide Location Analysis



Design and Manufacture of Intergenic Arrays



Content

Probe selection
- Intergenic regions

Coverage
- ~6300 IGR spots
- 95% of intergenic regions

Probe generation
- PCR products
- 95% success rate

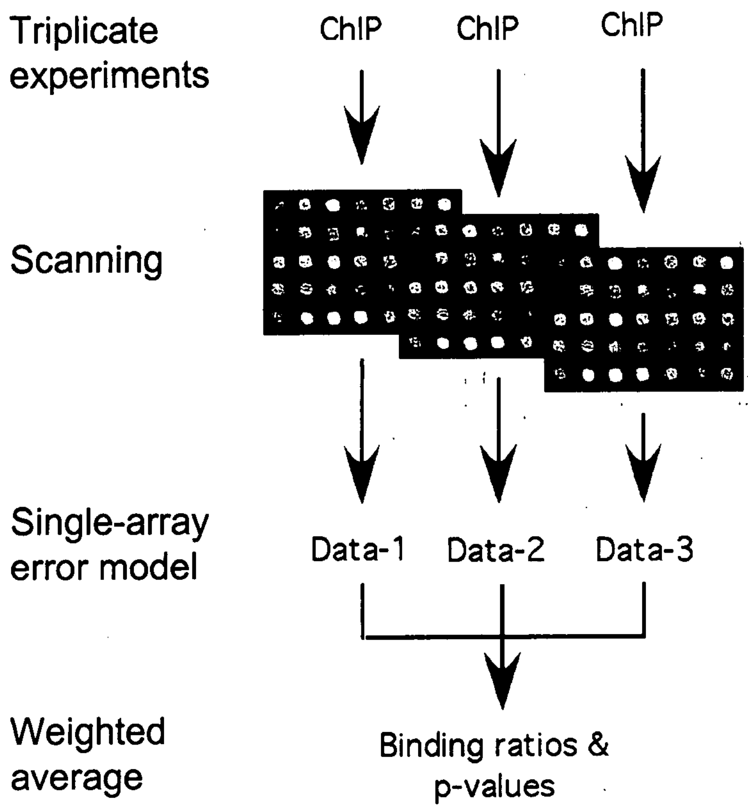
Controls
- Internal ORF regions
- Subgrid controls
- Serial dilution controls
- Exogenous DNA

Attachment Chemistry

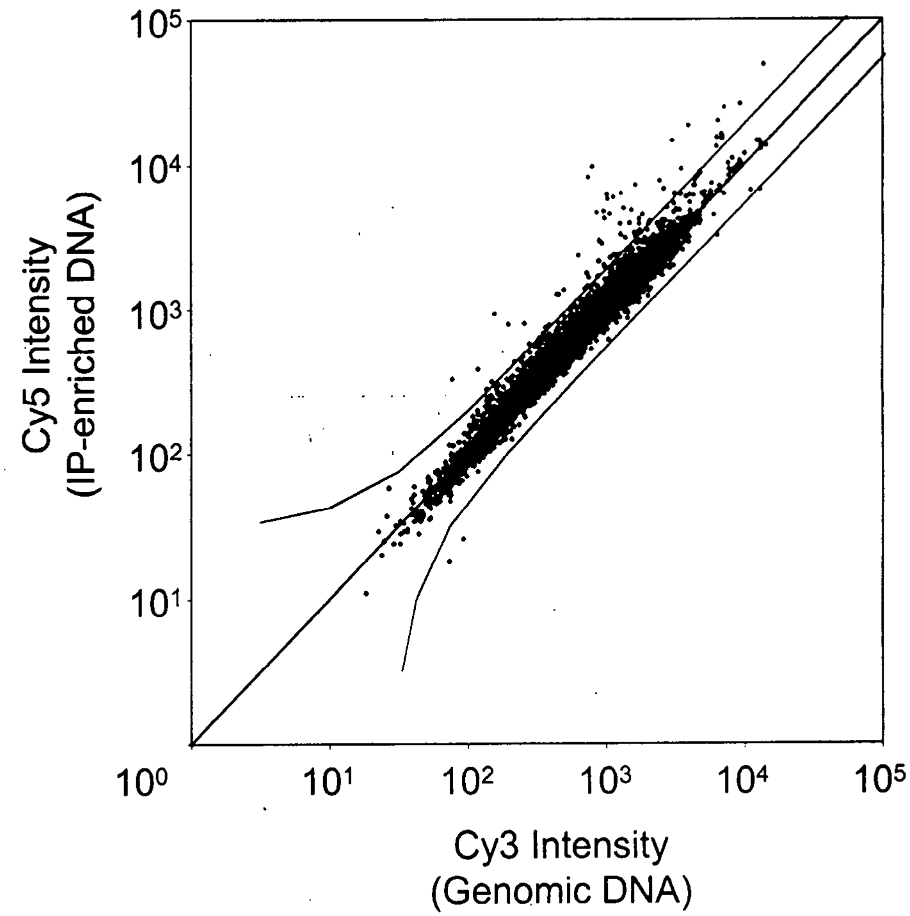
- Aminosilane

Analysis of Genome-wide Location Data

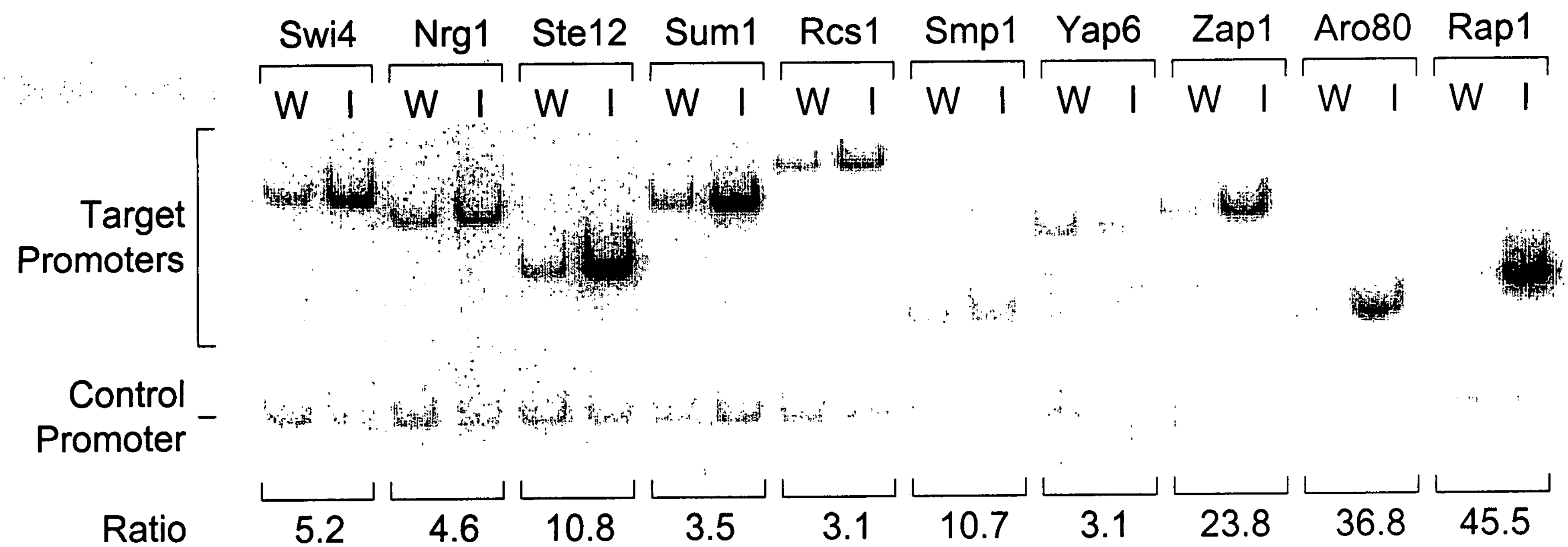
Data Generation



Error Model



False Positive and False Negative Rates



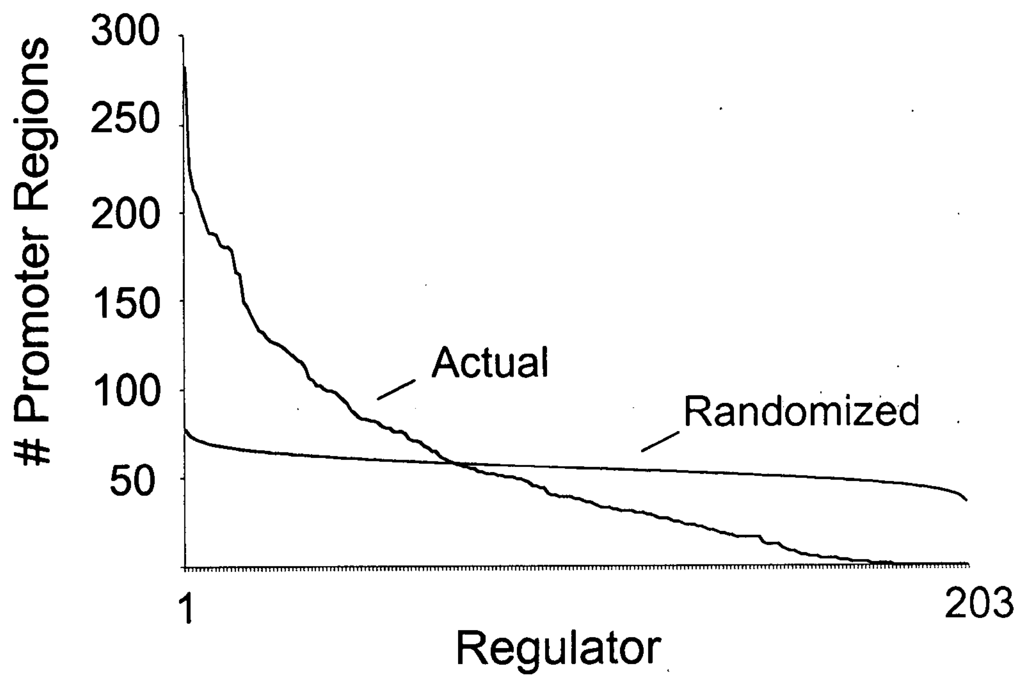
Using a p-value threshold of 0.001:

False positive rate ~6%
False negative rate ~30%

Summary of Genome-wide Binding Data

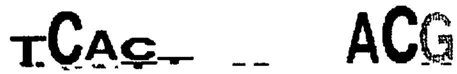
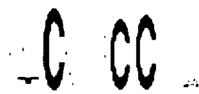
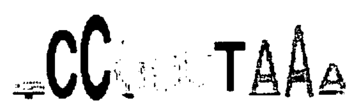
Promoter Regions Bound per Regulator

(High confidence: $P < 0.001$)

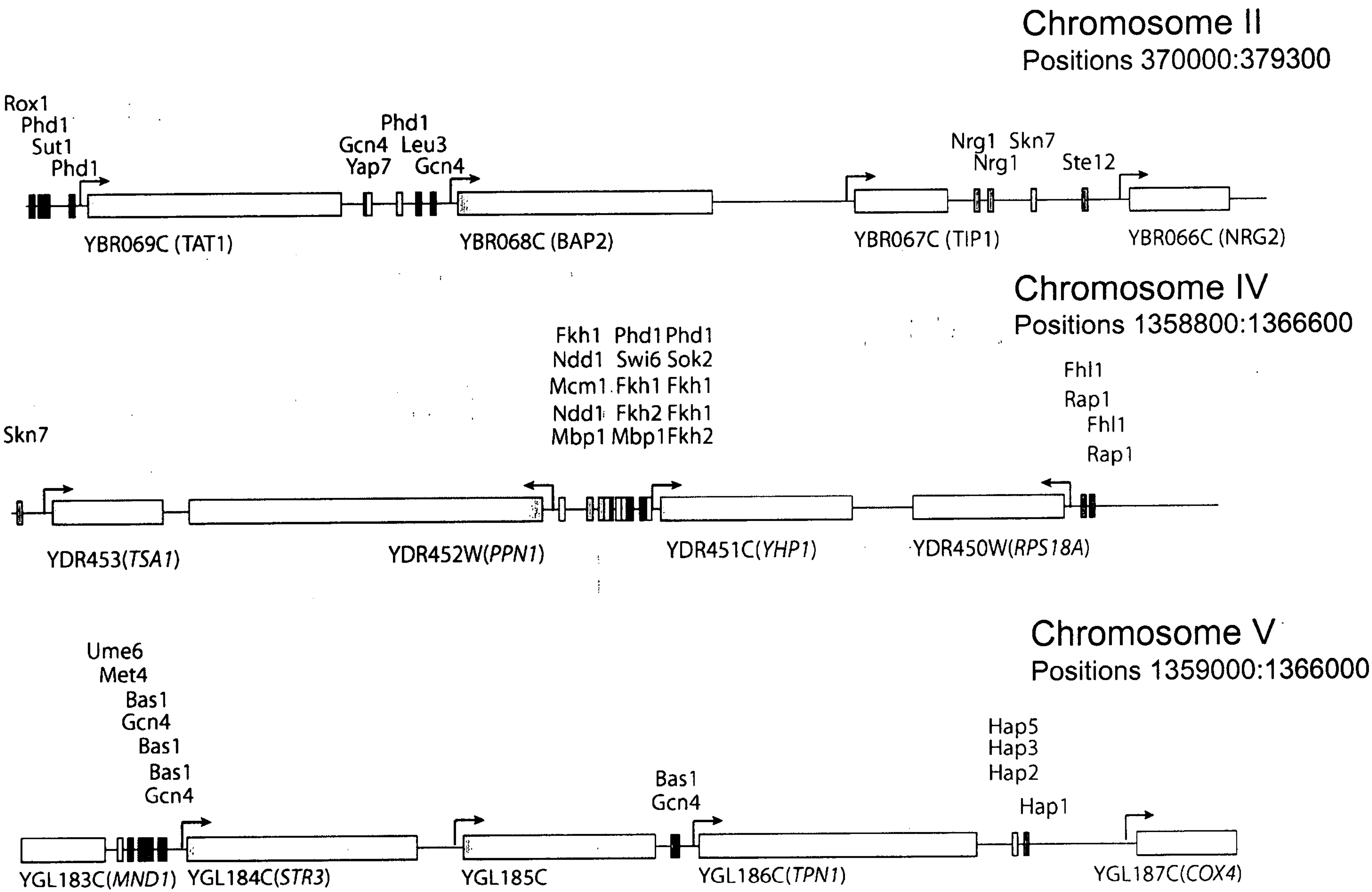


- Over 11,000 protein-DNA interactions
- Avg # promoter regions bound per regulator: 55
- Most abundant regulators bind to largest # promoters
- No promoter occupancy observed: 20 regulators

Regulator Binding Site Sequence Discovery

Re-discovered sequences:		Newly discovered sequences:	
Abf1		Aft2	
Cbf1		Cin5	
Gcn4		Phd1	
Hsf1		Rds1	
Leu3		Stb5	
Ste12		YDR026C	

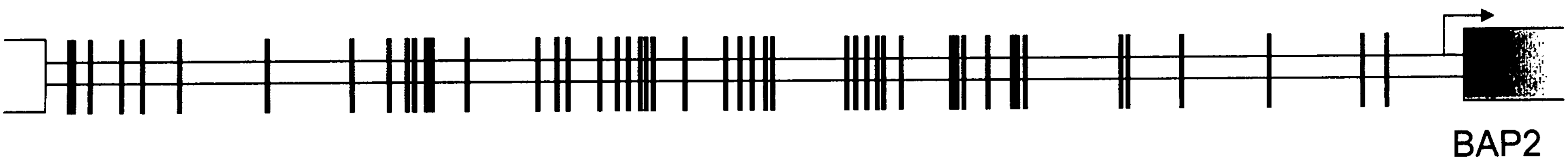
Samples of the Draft Transcriptional Regulatory Code



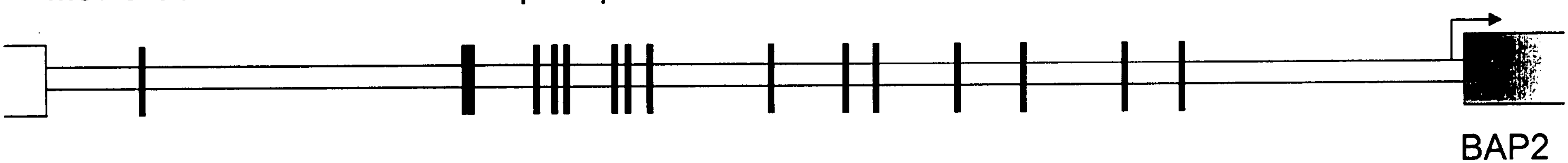
Influence of Data Types on Transcriptional Regulatory Code Discovery

Intergenic Promoter Region for BAP2

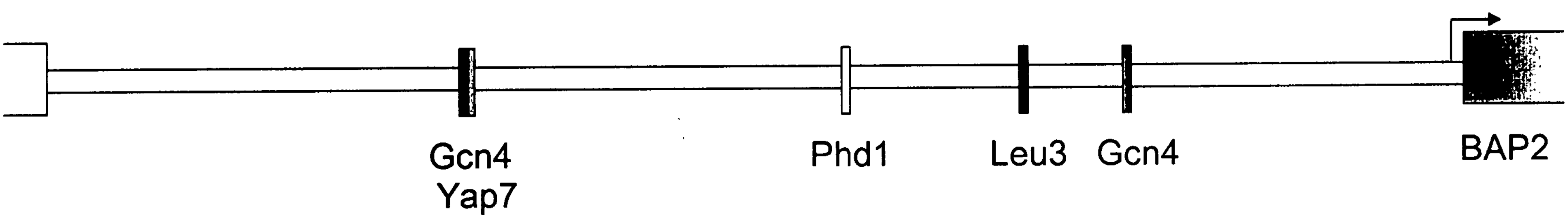
All Motifs



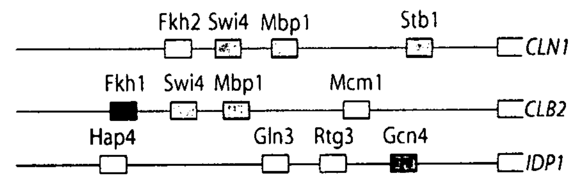
Motifs Conserved Across Multiple Species



Conserved Motifs AND Binding Information from Location Analysis



A Draft Transcriptional Regulatory Code for Yeast



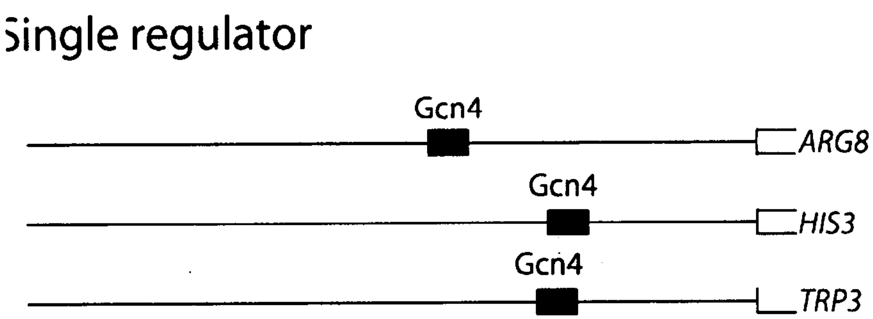
Introduction
- Concept
- History

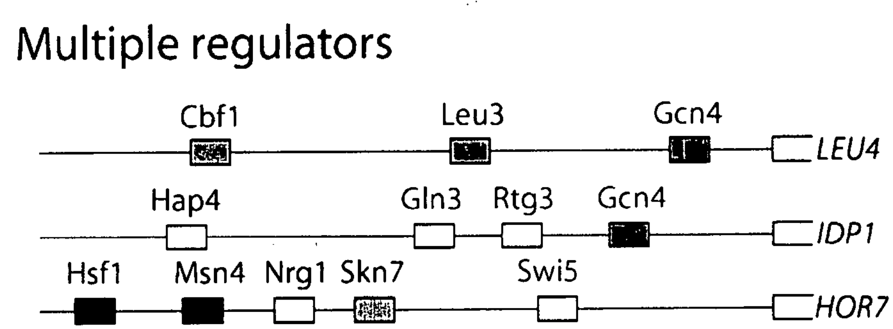
Data and Regulatory Code Assembly
- Genome-wide binding data
- DNA sequence specificities

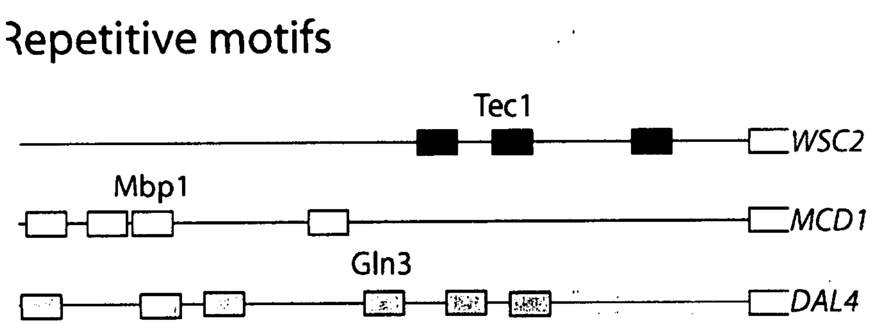
Promoter Architectures

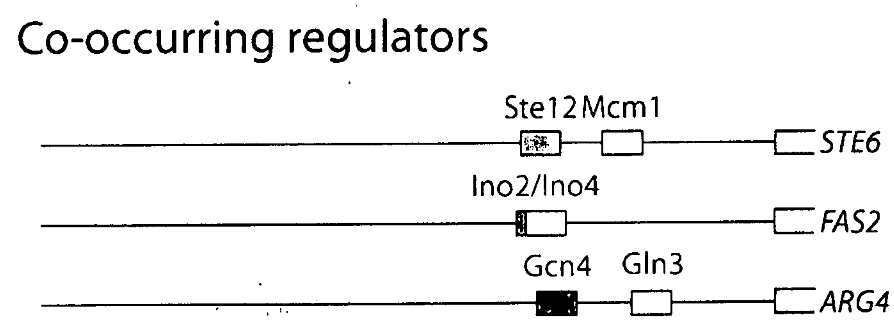
Environment-dependent Use of Code

Promoter Architectures

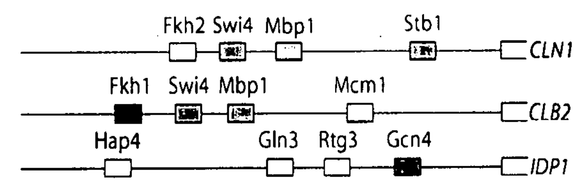








A Draft Transcriptional Regulatory Code for Yeast



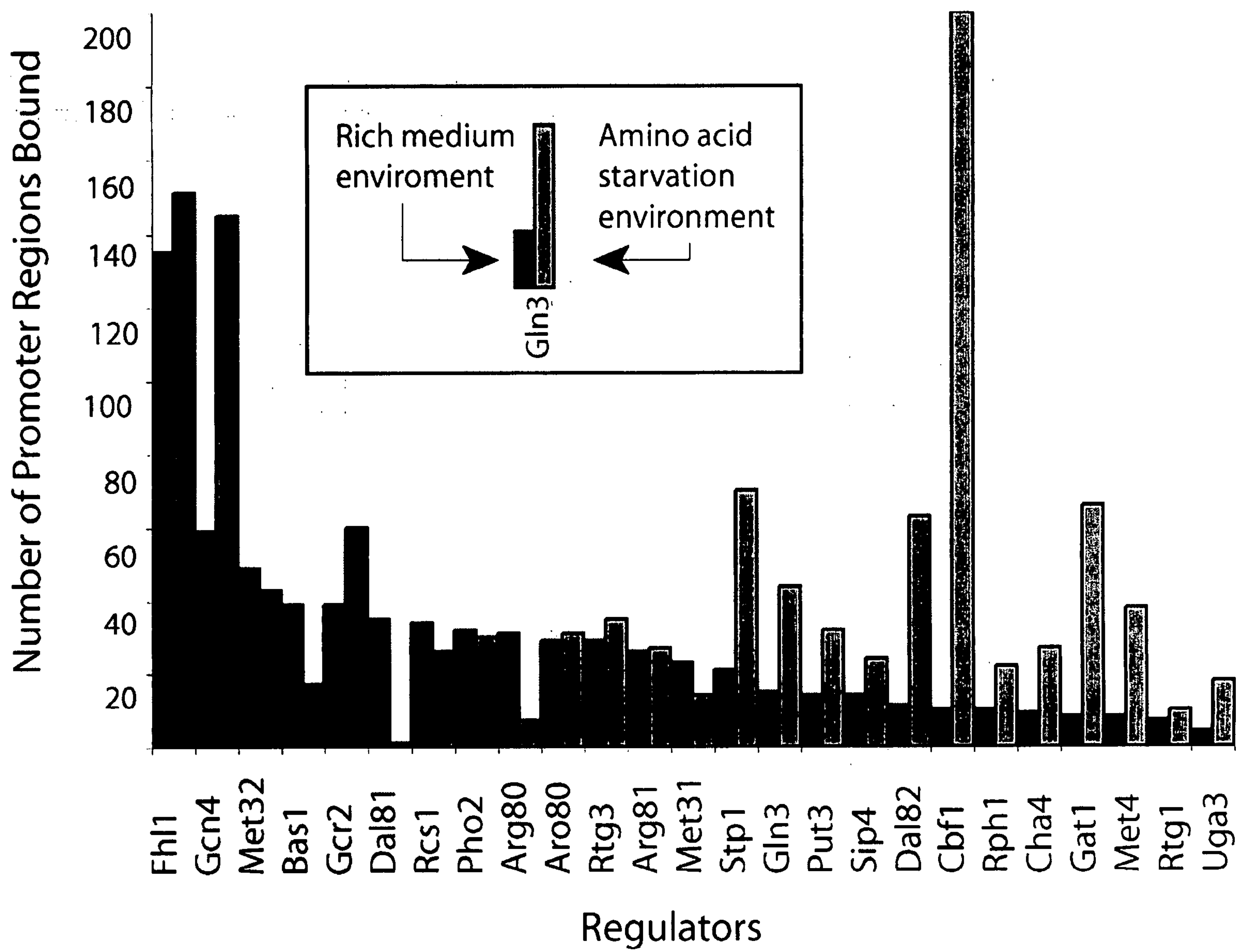
Introduction
- Concept
- History

Data and Regulatory Code Assembly
- Genome-wide binding data
- DNA sequence specificities

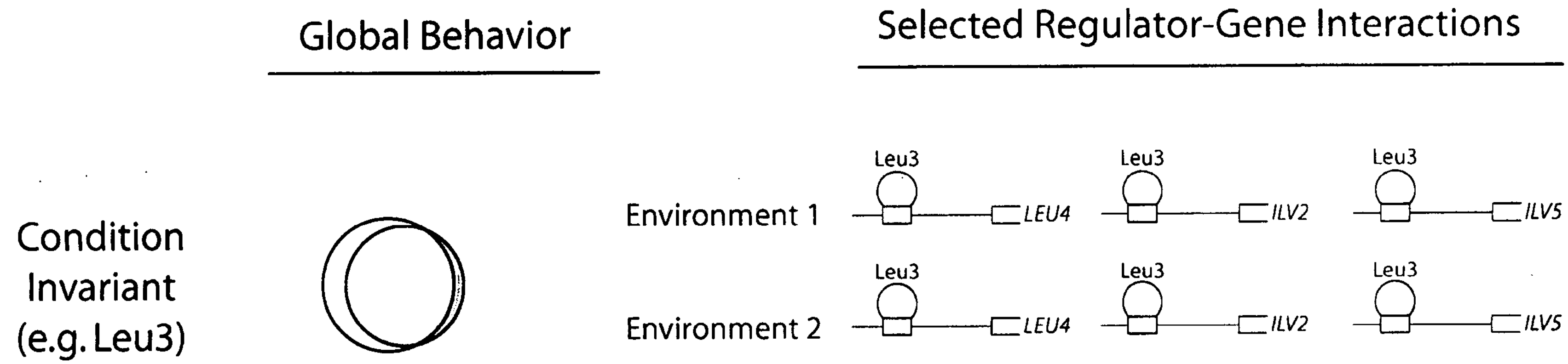
Promoter Architectures

Environment-dependent Use of Code

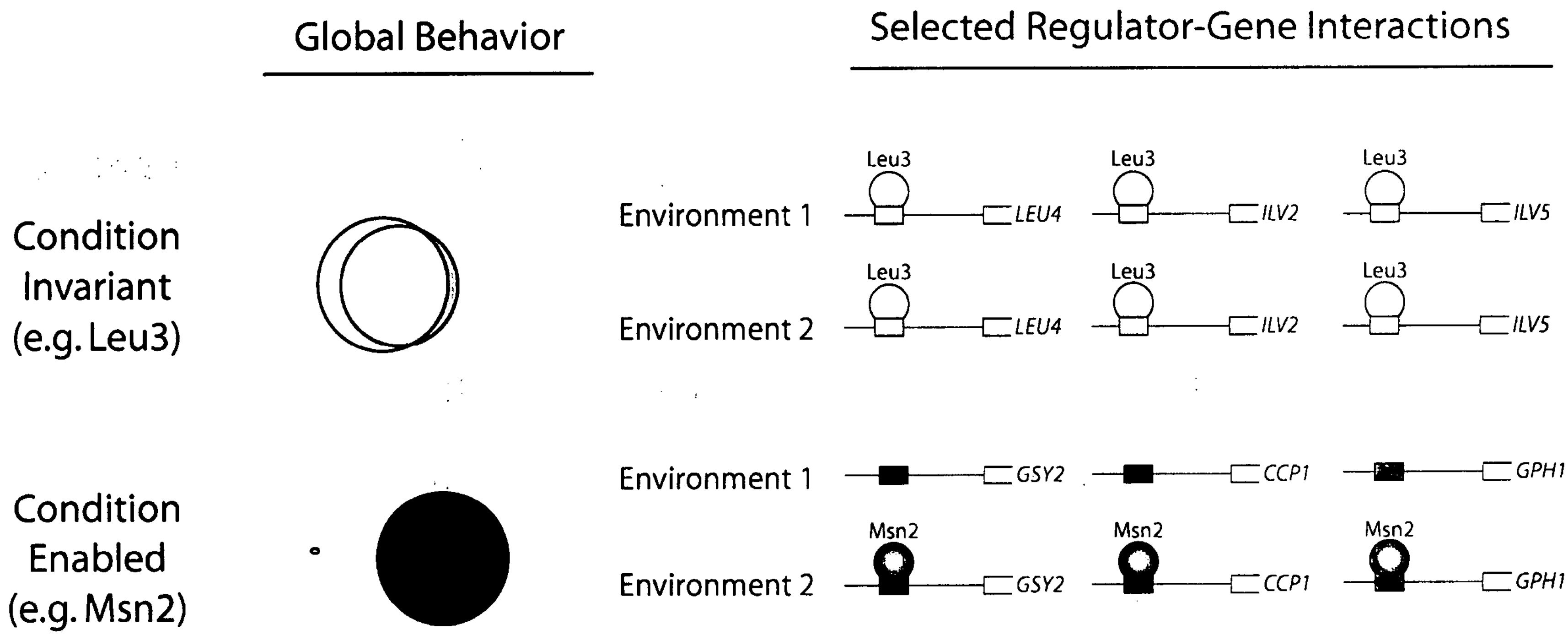
Environmental Effects on Genomic Binding



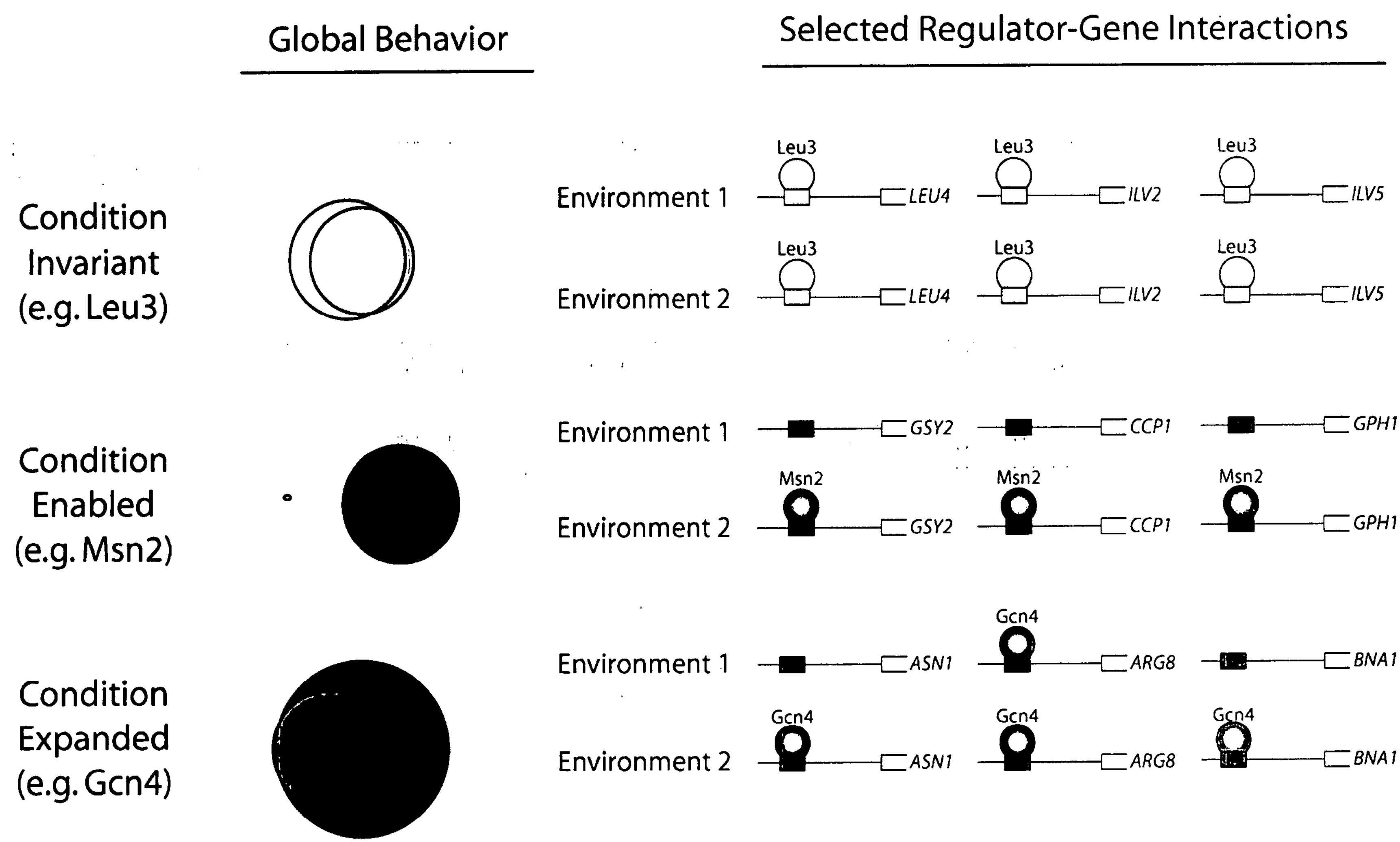
Environment-Specific Regulator Behaviors



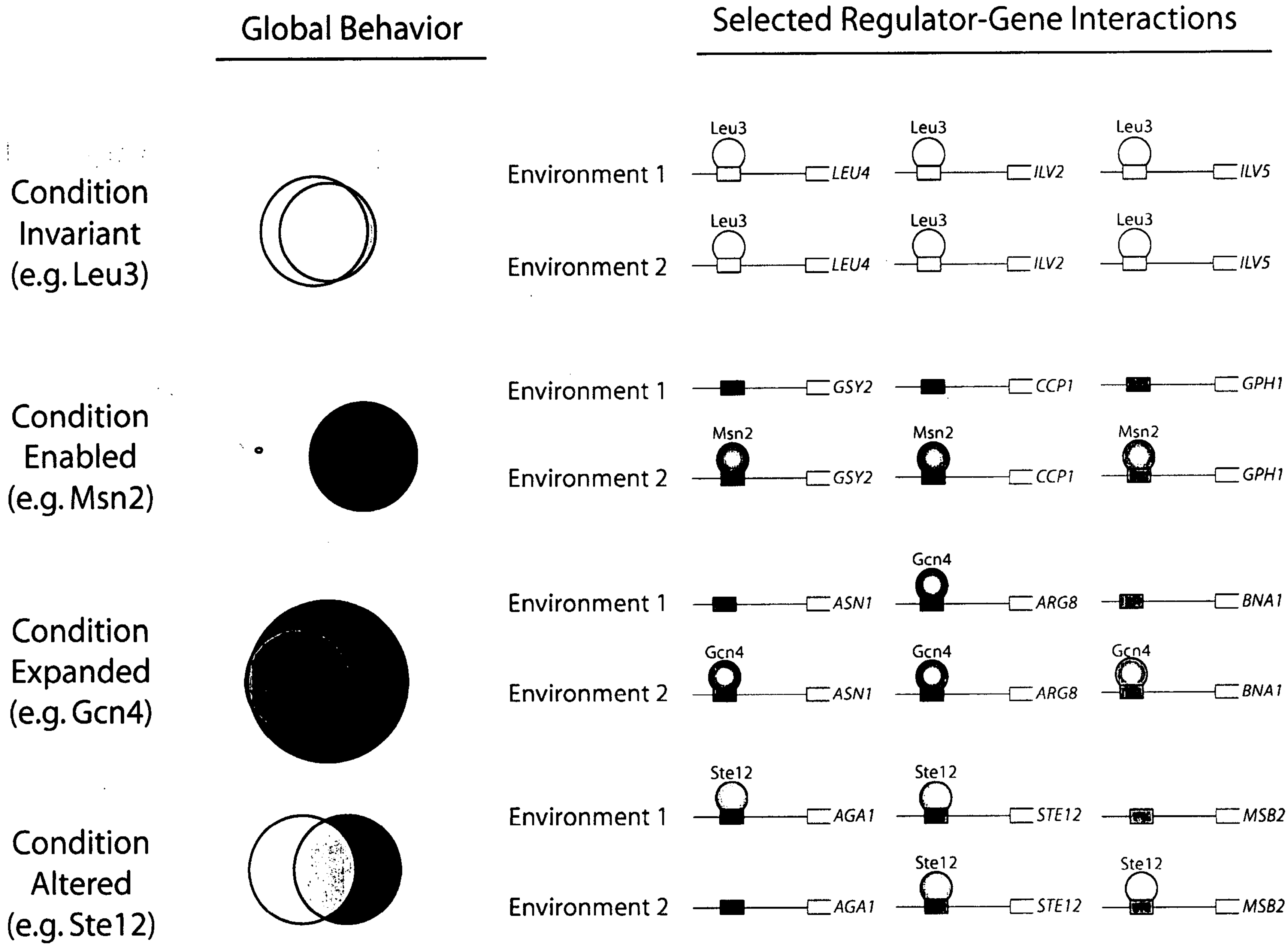
Environment-Specific Regulator Behaviors



Environment-Specific Regulator Behaviors



Environment-Specific Regulator Behaviors

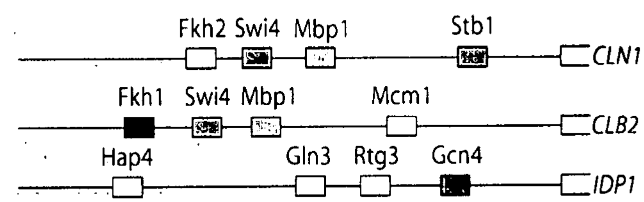


Challenges for Future Drafts

- Additional data
  - More DNA-binding regulators
  - Chromatin regulators
  - More environmental conditions
- Tests for regulatory mechanisms
  - Regulator abundance, modifications and translocation
  - Integration with expression information
- Refined computational tools

Regulation of Genome Expression in Health and Disease

Yeast Transcriptional Regulatory Code



First draft developed using combination of
- Genome-wide regulator binding data
- Sequence conservation data

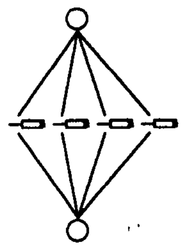
Ren et al. *Science* (2000)
Simon et al., *Cell* (2001)
Wyrick et al., *Science* (2001)
Lee et al., *Science* (2002)
Zeitlinger et al., *Cell* (2003)
Bar-Joseph et al., *Nature Biotech*nology (2003)

Environment-Specific Regulator Behaviors

Global Behavior

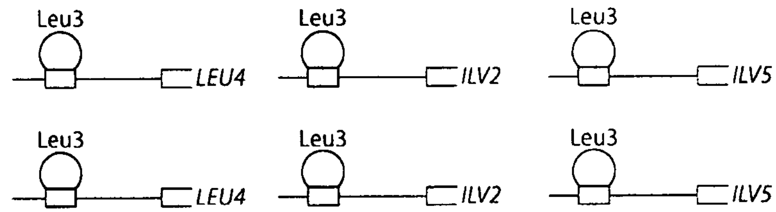
Selected Regulator-Gene Interactions

Condition Invariant (e.g. Leu3)

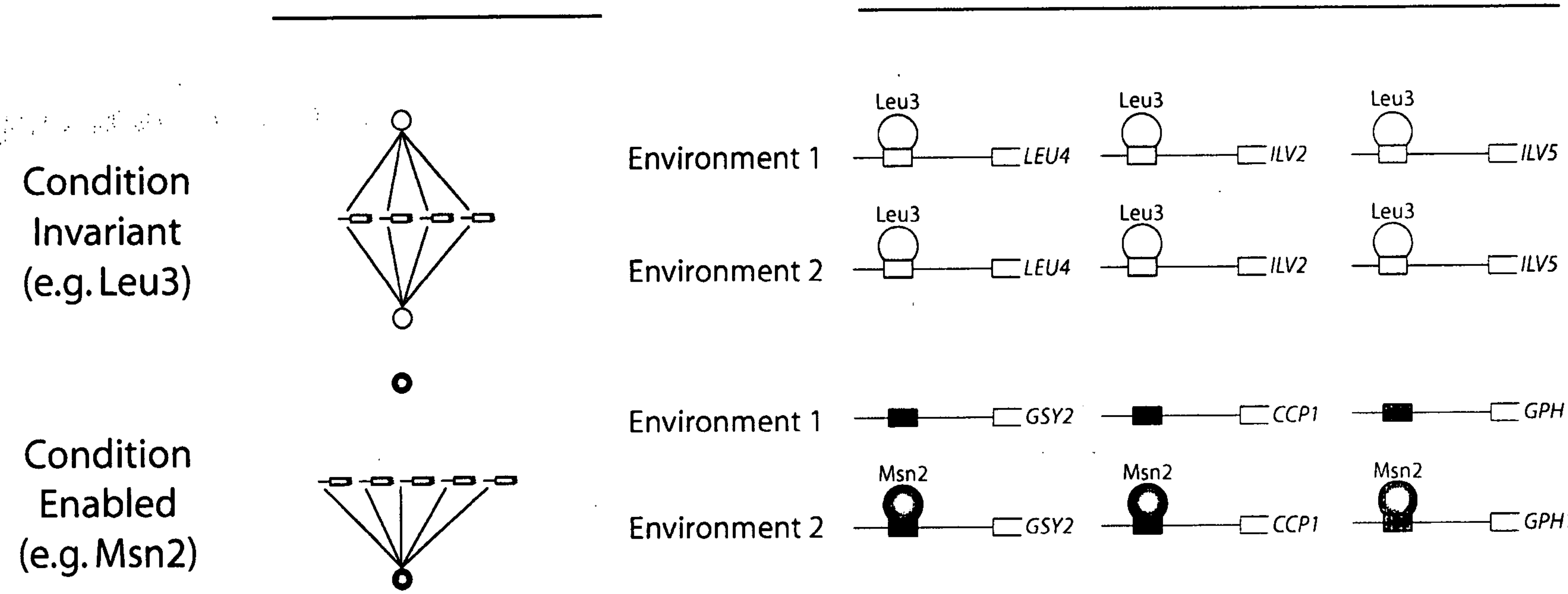


Environment 1

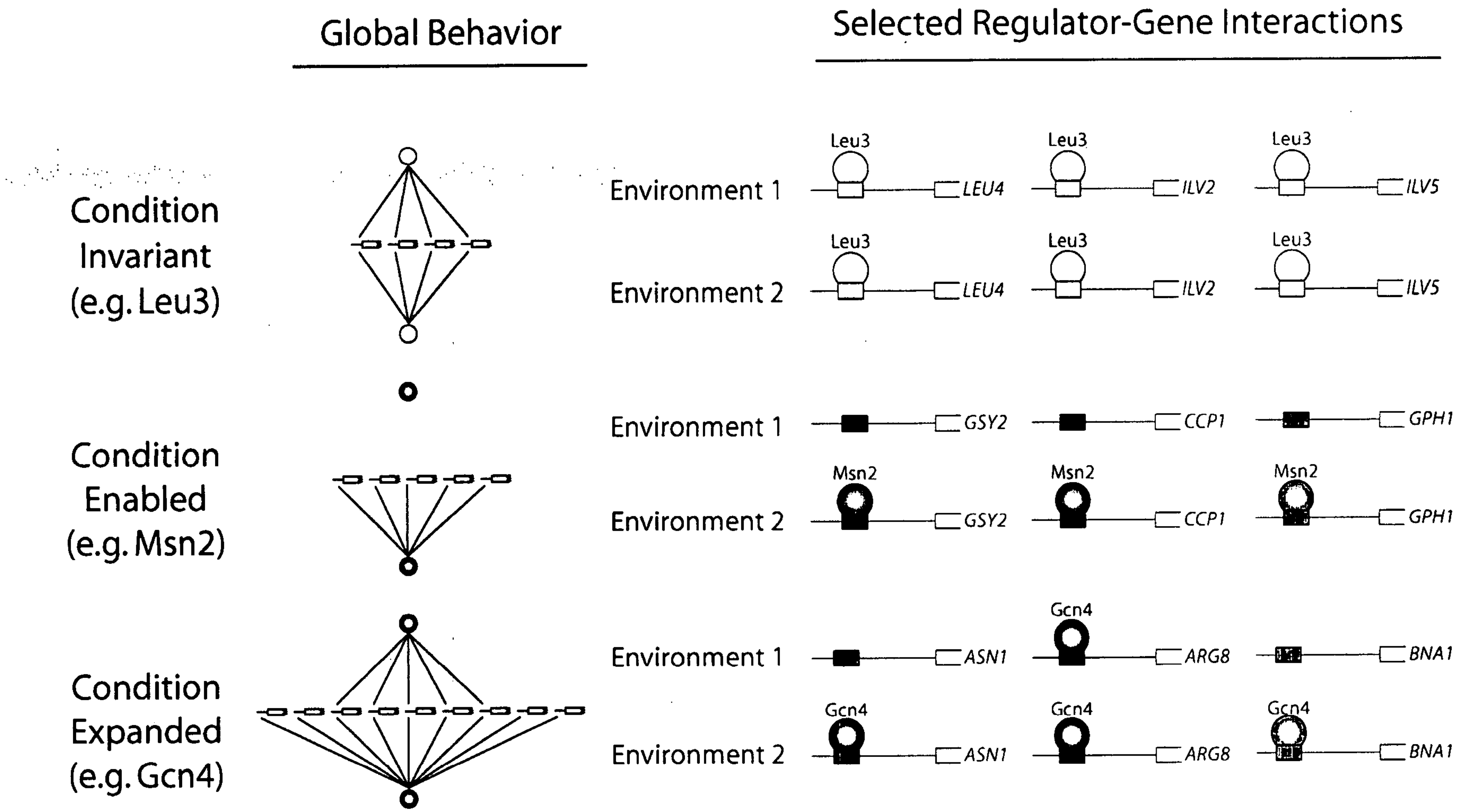
Environment 2



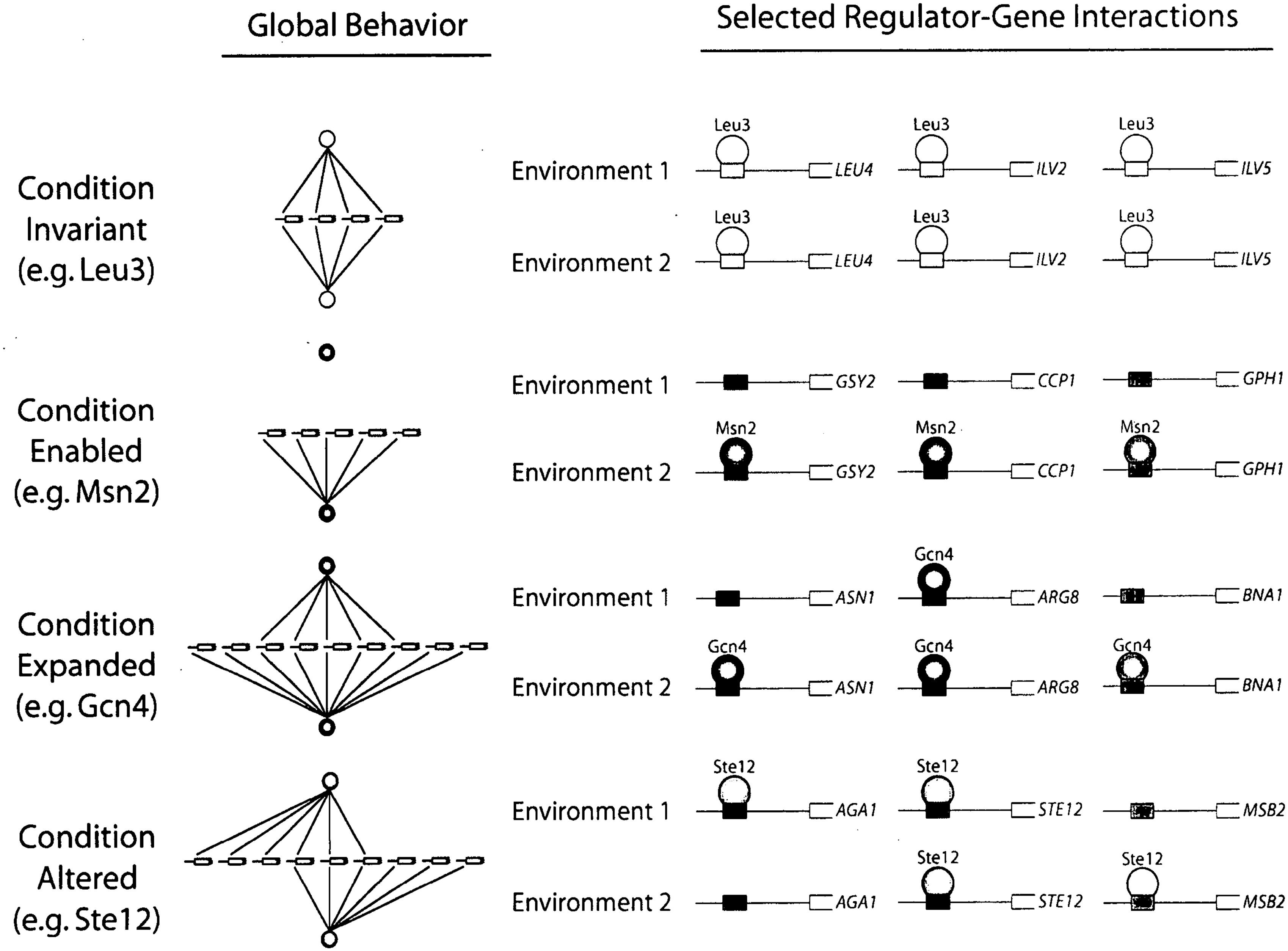
Environment-Specific Regulator Behaviors



Environment-Specific Regulator Behaviors



Environment-Specific Regulator Behaviors



Supplemental Application Data Sheet

Application Information

Application number::	Not Yet Assigned
Filing Date::	03/04/04
Application Type::	Provisional
Subject Matter::	Utility
Suggested Group Art Unit::	N/A
CD-ROM or CD-R?::	None
Sequence submission?::	None
Computer Readable Form (CRF)?::	No
Title::	TRANSCRIPTIONAL REGULATORY CODES OF EUKARYOTIC GENOMES AND METHODS THEREOF
Attorney Docket Number::	WIBL-P60-035
Request for Early Publication?::	No
Request for Non-Publication?::	No
Small Entity?::	No
Petition included?::	No
Secrecy Order in Parent Appl.?::	No

Applicant Information

Applicant Authority Type::	Inventor
Primary Citizenship Country::	US
Status::	Full Capacity
Given Name::	Christopher
Middle Name::	T.
Family Name::	Harbison
City of Residence::	Quincy
State or Province of Residence::	MA
Country of Residence::	US
Street of mailing address::	140 Quincy Avenue #23

City of mailing address::	Quincy
State or Province of mailing address::	MA
Postal or Zip Code of mailing address::	02169

Applicant Authority Type::	Inventor
Primary Citizenship Country::	US
Status::	Full Capacity
Given Name::	David
Middle Name::	B.
Family Name::	Gordon
City of Residence::	Somerville
State or Province of Residence::	MA
Country of Residence::	US
Street of mailing address::	103 Concord Avenue, Apt 4-D
City of mailing address::	Somerville
State or Province of mailing address::	MA
Postal or Zip Code of mailing address::	02143

Applicant Authority Type::	Inventor
Primary Citizenship Country::	US
Status::	Full Capacity
Given Name::	Richard
Middle Name::	A.
Family Name::	Young
City of Residence::	Weston
State or Province of Residence::	MA
Country of Residence::	US
Street of mailing address::	216 Highland Street
City of mailing address::	Weston
State or Province of mailing address::	MA
Postal or Zip Code of mailing address::	02193

Correspondence Information

Correspondence Customer Number::	28120

Representative Information

Representative Customer Number::	28120

Assignee Information

Assignee name::	Whitehead Institute for Biomedical Research
Street of mailing address::	Nine Cambridge Center
City of mailing address::	Cambridge
State or Province of mailing address::	MA
Postal or Zip Code of mailing address::	02142-1479